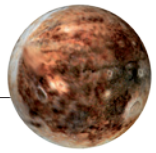
THIS WEEK



EDITORIALS

VACCINES Of past success and how to deliver on future promise **p.420**

WORLD VIEW How Europe can get the most for its Framework programme money **p.421**

PLUTO Carbon monoxide found around dwarf planet **p.423**

The long game

Graphene is not a miracle material, just a very promising one. It will take restraint and sustained interest to deliver its potential.

Is it possible for a field of science to move too fast? Perhaps. Those who work on the form of carbon known as graphene have seen it rocket from the next big thing to a miracle material in less time than it can take for a research paper to be accepted and published. Yet although the hunt is on for applications that can exploit graphene's remarkable properties, the work necessary to find out how it could best be harnessed remains incomplete.

This is one reason why many scientists who attended a meeting on the subject earlier this month were cautious about raising expectations too high. 'Graphene: The Road to Applications' was held in Cambridge, Massachusetts, on 11–13 May, and was hosted by Nature Publishing Group.

As with most 'overnight' success stories, the field of graphene research actually had a very slow start. Physicists have long been fascinated by this one-atom-thick planar crystal of carbon atoms, but until a few years ago most considered it a hypothetical system, useful for studying fundamental properties of matter but not able to exist in free form. It took a steady build-up of theoretical work and pioneering experiments over several decades before the research storm broke when it became possible to isolate graphene sheets in a laboratory.

The material certainly has an impressive set of properties. And there is no denying its tantalizing promise. As a thin membrane — flexible, strong and impermeable — graphene is an attractive platform on which to build devices. Add to that its high electron mobility and excellent optical and thermal properties, and the potential for applications seems endless. But the key word here is potential. The reality is that for many of graphene's widely promoted applications, it currently performs no better than existing materials and conventional approaches. For example, graphene performs relatively poorly as a transparent conductor in conventional displays, touch screens and photovoltaic cells. Neither is it a serious contender to replace silicon in standard electronics, because it does not work well as a digital switch, the essential function of a silicon transistor.

This should not necessarily be discouraging; the field is still very young. Those who talk up the potential of graphene's applications should remember that it typically takes any technology some 20 years to emerge from the lab and be commercialized — and even then it can succeed only with sustained effort and interest from researchers, industry and funding agencies (see *Nature* **469**, 14–16; 2011).

It is crucial to identify and acknowledge practical hurdles — and when it comes to fabricating and designing graphene devices, there are many. As was stressed many times at the Cambridge meeting, the single biggest obstacle is the lack of a way to reliably produce high-quality graphene sheets in large quantities. The quality and structure of the sheets can differ from batch to batch, which can make devices behave inconsistently.

Furthermore, although the pristine form of graphene has high electron mobility, meaning that devices based on it have the potential to run at high speeds, this changes dramatically when it interacts with its environment, as it does when attached to a substrate. The material will be valuable only if it can perform reliably when put to work.

Graphene stands out from other materials for its combination of superlative properties. In the right application, it could have an enormous impact. For example, although graphene does not perform well in digital electronics, there is much interest in using it to make high-frequency transistors for analogue applications — where it could be truly useful in wireless communication. Another important, if low-profile, role for graphene could be as a passive layer for heat management in electronic devices. Such targeting of graphene's potential will be vital to its success.

"It typically takes any technology some 20 years to emerge from the lab and be commercialized."

A few carefully chosen driver applications, although possibly not as obvious or exciting as some of the uses currently flagged, would give research on the material a better chance of holding on to wide interest in the long term — especially from funding agencies. Overnight success takes time to build. ■

Copy and paste

A slow university investigation into serious accusations of misconduct benefits no one.

As retractions go, it may not look like a big deal. Earlier this month, a statistics journal decided to pull a little-cited 2008 paper on the social networks of author–co-author relationships after it emerged that sections were plagiarized from textbooks and Wikipedia. The fact that this caused a wave of glee to ripple through the climate-change blogosphere takes some explaining.

Two of the paper's authors, Yasmin Said and Edward Wegman, both of George Mason University in Fairfax, Virginia, are also authors of an infamous 2006 report to Congress, co-written with statistician David Scott of Rice University in Houston, Texas. That report took aim at climatologist Michael Mann of Pennsylvania State University in University Park, suggesting that he was working in an isolated social network separated from "mainstream statisticians", and that he had such close ties with the rest of the field that truly independent peer review of his work was not possible. This report came to be known as the Wegman report, and has been frequently cited by climate-change sceptics.

This social-network analysis of Mann and his co-authors — with Mann's name removed — was cut down to an academic paper and published two years later in the journal *Computational Statistics &*

Data Analysis. It is this paper that the journal has decided to retract. So it seems likely that the plagiarism in the 2008 paper is also present in the 2006 Congress report. Still not look like a big deal?

That doubts about the 2006 report have resulted in concrete action is mainly down to the sterling work of an anonymous climate blogger called Deep Climate. His website first reported plagiarism in a different section of the congressional report in December 2009. One of those whose work was plagiarized is Raymond Bradley, director of the Climate System Research Center at the University of Massachusetts, Amherst. Ironically, Bradley was one of the co-authors of the climate reconstructions criticized by the Wegman report. Bradley, alerted by Deep Climate, complained to George Mason University on 5 March last year.

Wegman has blamed a graduate student for the plagiarism. Daniel Walsch, spokesperson for George Mason University, says that an internal review of the matter began in the autumn. He cannot estimate when that review will be complete, and, until it is, he says, the university regards it as a "personnel matter" and will not comment further. He adds that the review is still in the "inquiry" phase to ascertain whether a full investigation should be held. "Whether it is fast or slow is not as important as it being thorough and fair," says Walsch.

The fact that 14 months have passed since Bradley's complaint without it being resolved is disheartening but not unusual. An examination of George Mason University's misconduct policies suggests that investigations should be resolved within a year of the initial complaint, including time for an appeal by the faculty member in question. According to the university's own timeline, the initial inquiry should have been complete within 12 weeks of the initial complaint — in May 2010. But there are loopholes galore for extensions, and, like many universities, George Mason seems content to drag its feet.

Long misconduct investigations do not serve anyone, except perhaps university public-relations departments that might hope everyone will have forgotten about a case by the time it wraps up. But in cases such as Wegman's, in which the work in question has been cited in policy debates, there is good reason for haste. Policy informed by rotten research is likely to have its own soft spots. Those who have been wronged deserve resolution of the matter. And one can hardly suppose that those who have been wrongfully accused enjoy living under a cloud for months.

So, what incentives do universities have to pick up the pace? Agencies such as the US Office of Research Integrity and ethics offices at funding bodies should take universities to task for slow investigations and demand adherence to the schedules listed in university policies. However, the agencies themselves haven't exactly been models of swift justice. The most recent annual report from the Office of Research Integrity — for 2008 — reported that the cases closed in that year spent a mean of 14.1 months at the agency. Perhaps it should fall to accreditation agencies to push for speedy investigations. Tom Benberg, vice-president of the Commission on Colleges of the Southern Association of Colleges and Schools — the agency that accredits George Mason University — says that his agency might investigate if the university repeatedly ignored its own policies on the timing of misconduct inquiries. To get the ball rolling, he says, someone would have to file a well-documented complaint.

"If the work in question has been cited in policy debates, there is good reason for haste."

Even if funding and accreditation agencies fail to apply pressure, universities should take the initiative to move investigations along as speedily as possible while allowing time for due process. Once an investigation is complete, the institution should be as transparent as it can about what happened. Especially when public funds are involved, or at public universities, the taxpayer has a right to know what happened when papers are retracted — even if the faculty member in question is eventually exonerated. This tidies the scientific record, clears the air and kicks the legs out from under any conspiracy theories. Over to you, George Mason University. ■

Modern heroes

The great achievements of vaccines are not consigned to the past.

It is easy to see the heroic age of vaccines as one that ended decades ago. The Salk polio vaccine, after all, which swiftly and visibly transformed the disease into a distant memory in the developed world, was introduced in 1955. And the smallpox eradication campaign led by the World Health Organization had, by the late 1970s, reduced the virus from a killer of millions of people a year to a prisoner of biosafety labs. These were monumental feats, but the best could be still to come.

This week *Nature* explores the undiminished promise of vaccines, and the factors that threaten it — complacency, funding shortages and the unease that vaccines provoke in so many people.

Worldwide, up to one-third of all deaths of children under five result from diarrhoea and pneumonia. In the past ten years or so, vaccines against the microorganisms that cause many of these cases have become a standard part of the childhood regimen in the developed world. If they could be made available worldwide, the lives of hundreds of thousands of children could be saved each year.

Research efforts are adding to the promise. Together, AIDS, malaria and tuberculosis kill more people each year than smallpox did when the global campaign to eradicate it began in 1967. The search for vaccines for all three diseases has been long and frustrating, but a Perspective on page 463 describes how new technologies are reviving it.

There is no room for complacency. The global campaign to eradicate polio made stunning progress from 1988 to the end of the twentieth century, reducing worldwide incidence by 99%. But the disease continues to smoulder in Pakistan, India, Afghanistan and Nigeria, where vaccinators have struggled with turmoil and corruption, high transmission rates and suspicion about the vaccine itself (see pages 427 and 446). Similarly, a long vaccination campaign against measles has reduced the global death toll from more than 2.5 million a year in 1980 to fewer than 200,000 today. But vaccination rates are still below 80% in much of Africa and India, and funds pledged to the global measles initiative have fallen. Some people think that the disease is poised to surge again in the developing world (see page 434). Europe has already seen outbreaks, in part because vaccination rates dipped after the combined measles, mumps and rubella (MMR) vaccine was falsely linked to autism.

Vaccines can become victims of their own success. In the developed world, for example, vaccination has already reduced measles to a rarity, which makes an 'informed' choice to shun the vaccine seem risk free. Even doctors and nurses can fall prey to this reasoning. They have a disproportionate influence over whether parents vaccinate their children, and when they lose sight of the overwhelming ratio of benefit to risk for most vaccines, they can amplify public fears (see page 443). Back in the 1950s, '60s, and '70s, when vaccines offered protection against clear and present menaces, it was easier to accept their small risk of harm.

Designing a cheap, effective vaccine against the more complex major killers of today is a harder task, and people everywhere are quicker to question the official line, on vaccines as on everything else. But the promise for vaccines to transform global health is as bright as ever, and funders and public-health experts must continue their heroic support for research, global vaccination efforts and communication strategies to win over the doubters. ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv

Can Europe build a framework for success?

The European Framework programme, one of the world's largest science funders, has improved its reputation. Not by enough, says **Colin Macilwain**.

Last week saw the passing of European scientists' best chance to help shape their single largest source of funding: the European Union's Framework programme. Their efforts are unlikely to bequeath a system fit to tackle the continent's pressing needs.

The consultation exercise that ended on 20 May is part of a convoluted process to design the Framework programme that involves lobbyists from science and industry, the European Commission, the European Parliament and the member states. It may prove unable to give the Framework the sharp focus it so badly needs. Too many trade-offs and compromises will, as always, see incremental change to a programme that is crying out for a radical overhaul.

The Framework is often derided for supporting second-rate projects, but much of what it now supports is excellent. This year, it will hand out more money for research than any other programme in the world, with the exception of NASA and the US National Institutes of Health.

James Heckman, a Nobel-prizewinning US economist, for example, will soon be spending much of his time in Ireland, having won a grant from the European Research Council (ERC) — part of the Framework — to study health economics at University College Dublin. This doesn't reverse a century of academic brain-drain west across the Atlantic, but it is a start.

The main problem remains the deficiencies in the largest part of the Framework programme, which supports targeted research projects carried out by partners from at least three nations (see *Nature* **464,** 349; 2010). Scientists have long grumbled that many of these projects are awarded large (sometimes very large) grants not because of their research excellence, but because political winds blow in their direction.

The Eighth Framework Programme (FP8) will run from 2014 to 2020, and is expected to cost around €70 billion (US$98 billion). So what do Europe's scientists hope to get from it? Judging by their published submissions (see http://go.nature.com/oipiho), they want — surprise, surprise — more money, more emphasis on excellence, simplified procedures, and a support structure for major new facilities.

Let's look at each of these objectives in turn. The outlook for increased funds is forlorn. The seven-year budget is being planned at the worst possible time and will be set in 2012/13 — the time of maximum austerity for most of the 27 national governments footing the bill.

'Excellence' in this context means primarily the ERC, which was set up as part of FP7 to distribute grants purely on the basis of scientific merit. But questions remain about ERC governance; it has no director-general, and its expansion will be resisted by the large number of member states who can't really compete for ERC grants.

EUROPE SHOULD FOCUS MORE ON BACKING THE BEST PEOPLE WITH THE BEST IDEAS IN ALL BRANCHES OF SCIENCE.

Advocates of 'simplification' often call for a trust-based system with less paperwork and auditing. But not everyone in Brussels agrees that researchers can always be trusted. One reason for this is that projects need multiple partners to win funding, yet, once handed the money, not all of those partners pull their weight. Add vivid memories of fraud allegations against former research commissioner Edith Cresson, and it's no wonder the commission's auditors want to keep a keen eye on the Framework programme.

There seems to be wide agreement that the programme could help with infrastructure. But, at present, there is no established mechanism to build and run European facilities, resulting in tricky negotiations between up to 27 nations for every proposed facility — and a recurrent impasse between the technologically advanced and less-advanced nations on who should host them.

These, then, are what scientists want from FP8. The commission, alas, seeks something else.

The first thing it wants is 'innovation', the watchword for Máire Geoghegan-Quinn since she took over the research directorate — now the research and innovation directorate — early last year. Like many politicians, she seems hazy on the distinction between research and innovation, and reluctant to acknowledge limits in the potential of state actions to stimulate the latter. The commission's other goal is to align research programmes more closely with 'cohesion' — Eurojargon for helping poor countries on the European Union's periphery to catch up with its Germanic core.

Such an alignment could pull Framework money away from excellence and the expansion of the ERC. This fight will be at the heart of the coming tussle over FP8. It is a fight that rich member states are likely to win, and so keep research funding largely separate from cohesion goals. That will please well-resourced scientists in places such as the United Kingdom and Germany, but anger their colleagues to the south and east.

A future strategy in Europe marked by continuity rather than change will be good enough for most grantees — but not good enough for those, including Geoghegan-Quinn, who argue that Europe faces a massive competitiveness crisis that can only be averted by a step-change in its innovative capacity.

Geoghegan-Quinn is right to demand drastic change, but wrong on the direction it should take. Instead of chasing the impossible goal of an 'Innovation Union' by broadening the Framework's reach, Europe should look to the model of the US National Science Foundation, further develop the ERC, and focus more on backing the best people with the best ideas in engineering, the humanities and all branches of science. ■

➲ NATURE.COM
Discuss this article online at:
go.nature.com/fhcj3e

Colin Macilwain *is a contributing correspondent with* Nature.
e-mail: colinmacilwain@googlemail.com

RESEARCH HIGHLIGHTS

Selections from the scientific literature

COGNITIVE NEUROSCIENCE

Scenes deciphered from spaces

When a person views a scene, be it a city street or grassy hills, the brain's parahippocampal place area (PPA) processes it mainly on the basis of its spatial characteristics, not its contents.

That's the finding of Dwight Kravitz and his colleagues at the National Institute of Mental Health in Bethesda, Maryland, who used functional magnetic resonance imaging to scan the brains of volunteers as they viewed 96 scenes ranging from open, natural environments to enclosed rooms. Scenes that were spatially similar — such as those depicting either open or closed environments — elicited similar PPA responses. However, scenes with the same kind of content — for example, man-made features — did not.

Another set of volunteers categorized the scenes on the basis of spatial and non-spatial features. Their ratings of the spatial features correlated with the PPA patterns; the non-spatial ratings did not.

J. Neurosci. 31, **7322–7333 (2011)**

KLINGER & LUO, CARNEGIE MUSEUM OF NATURAL HISTORY

PALAEONTOLOGY

Stronger smell, bigger brain

Mammals may owe their large brains to the development of more acute senses, such as smell and touch, in their extinct ancestors.

Timothy Rowe at the University of Texas in Austin and his co-authors made computed tomography scans of the intact fossil skulls of two species that preceded the first mammals. Compared with those of its predecessors, the brain of *Morganucodon oehleri* — which roamed the Earth some 200 million years ago — was larger relative to the size of its body. Much of the difference is attributable to the growth of brain areas involved in sensing and processing smell and touch, as well as movement coordination.

Another mammalian ancestor from the same period, *Hadrocodium wui* (**pictured**), also showed brain growth, particularly in regions attuned to smell. The authors suggest that an improved sense of smell might have laid the neural groundwork for the ability to deal with different types of environmental information.

Science 332, **955–957 (2011)**

For a longer story on this research, see go.nature.com/ngdo6o

ECOLOGY

Parasites make their hosts hide

Many parasites modify their host's behaviour, upping the host's risk of being ingested by a predator, and thus hastening the next stage of the parasites' life cycle. But, say Lucile Dianne and her colleagues at the University of Bourgogne in Dijon, France, one parasite can also do the opposite, making its victims predator-averse.

Pomphorhynchus laevis worms live in the amphipod *Gammarus pulex* (**pictured**) until they reach a stage at which they can infect the fishes that consume the amphipods. At this point, the amphipods grow increasingly reckless. But the researchers found that, in an experimental set-up, *G. pulex* infected with parasites that are at an earlier developmental stage spend more time hiding than uninfected individuals. This hiding behaviour protects the creature against fish predation, allowing time for the worm parasites to prepare for their next victim.

A. HYDE/NATURE PICTURE LIBRARY

Evolution **doi:10.1111/j.1558-5646.2011.01330.x (2011)**

GEOPHYSICS

Glacial biography of Greenland

Greenland's three largest glaciers lost an enormous amount of ice during the past decade, but each has quite different prospects for long-term stability. Ian Howat at Ohio State University in Columbus and his colleagues combined remote-sensing data with meteorological modelling to estimate the amounts of ice gained or lost from the glaciers each month from 2000 to 2010.

Despite a drastic retreat between 2004 and 2006, Helheim glacier managed to gain a small amount of mass by the end of the period. The Jakobshavn glacier, however, is shedding ice ever faster. Meanwhile, at Kangerdlugssuaq, mass loss sped up but has since returned to the 2000 rate. These differences, the researchers say, show that simply extrapolating from recent changes is not a reliable way of predicting future ice loss.

Geophys. Res. Lett. **doi:10.1029/2011GL047565 (2011)**

METABOLIC ENGINEERING

Bacterial chemical factories

An engineered strain of the bacterium *Escherichia coli* can turn sugars such as glucose into a commercially important chemical.

Stephen Van Dien of Genomatica in San Diego, California, and his colleagues used computer models to design a biochemical pathway that converts common *E. coli* metabolites into 1,4-butanediol (BDO), which is used to make various plastics. For each step of the process, the researchers then inserted into the *E. coli* genes that normally encode enzymes in other bacterial species. They also tinkered with the bacterium's metabolism to drive energy and carbon molecules through this new pathway. The top-producing strain pumped out 18 grams of BDO per litre — a yield that would need to increase three- to fivefold to be commercially viable.

The bacteria could one day offer a low-cost way to make BDO from renewable products, rather than petroleum.

Nature Chem. Biol. **doi:10.1038/nchembio.580 (2011)**

NEUROSCIENCE

Rapid recovery from stroke

Healthy parts of the brain can compensate for areas damaged by a stroke. Some of these changes are probably a result of the 'unmasking' of previously unused pathways rather than active rewiring.

Timothy Murphy and his colleagues at the University of British Columbia in Vancouver, Canada, induced small strokes in mice and then monitored their sensory brain activity for up to two hours while stimulating the animals' forepaws. If the stroke occurred in the right hemisphere (controlling the left paw), for example, sensory activity was higher in the left hemisphere when the left or right paw was stimulated, compared with sensory activity prior to the stroke.

Some of the changes occurred as early as 30 minutes after the stroke, too soon to have resulted from circuit rewiring. The authors suggest that stroke unleashes electrical signals that remove the inhibition that normally blocks certain existing pathways.

Proc. Natl Acad. Sci. USA **doi:10.1073/pnas.1101914108 (2011)**

GENETICS

Genetics of malaria severity

The mutation of just one DNA base in a gene is associated with protection against severe childhood malaria.

The *FAS* gene encodes a cell-surface protein called CD95, which promotes programmed cell death and has previously been implicated in severe malaria. Kathrin Schuldt at the Bernhard Nocht Institute for Tropical Medicine in Hamburg, Germany, and her colleagues looked for variants of the gene associated with severe disease in nearly 1,200 infected children in Ghana, West Africa. They found that a single nucleotide substitution in the gene's promoter — a regulatory region — corresponded to increased levels of CD95 and reduced the risk of severe malaria by 29%. They confirmed this finding in another 1,412 cases.

The authors speculate that boosting CD95 promotes immune-cell suicide, preventing an excessive immune response and lessening the disease's severity.

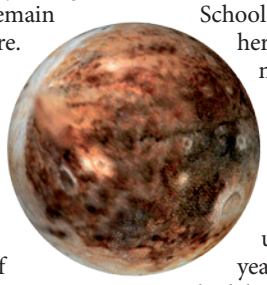
PLoS Genet. 7, **e1002066 (2011)**

PLANETARY SCIENCE

Carbon monoxide on Pluto

Pluto's nitrogen-rich atmosphere may contain a thin layer of carbon monoxide, according to Emmanuel Lellouch at the Observatory of Paris and his colleagues.

The surface of the dwarf planet (**pictured**) is known to be covered mostly by nitrogen ice, but questions remain about its atmosphere. The researchers confirmed the presence of gaseous methane and found evidence for carbon monoxide at concentrations of about 0.6% and 0.05%, respectively. They observed the absorption of light that is characteristic of the molecules using a European Southern Observatory telescope in Cerro Paranal, Chile.



F. SAURER/SCIENCE PHOTO LIBRARY

On the basis of modelling work, the team proposes that the carbon monoxide is concentrated either in a thin layer in Pluto's atmosphere, or, less likely, in pure patches covering 0.2–1.2% of the surface.

Astron. Astrophys. 530, **L4 (2011)**

CANCER

Radiation's double whammy

Radiation is known to cause cancer by damaging DNA, but may also induce other molecular changes in the surrounding tissue that accelerate tumour growth.

Mary Helen Barcellos-Hoff at the New York University School of Medicine and her colleagues exposed mice to radiation and then inserted cancer-prone tissue into their mammary glands and those of untreated mice. One year later, tumours had developed in all of the irradiated mice, but in only 69% of non-irradiated mice. The tumours also grew faster in the irradiated animals, and more of them were oestrogen-receptor negative, a marker of aggressive breast cancer. Furthermore, the radiation activated a protein in the surrounding tissue called TGF-β, accelerating cancer development.

The findings might help to explain why women who receive radiotherapy treatment for childhood cancer are at greater risk from early-onset breast cancer later in life.

Cancer Cell 19, **640–651 (2011)**

COMMUNITY CHOICE

The most viewed papers in science

ENGINEERING

See-through solar cells

HIGHLY READ on apl.aip.org in April

Light striking the windows of houses and skyscrapers could one day be harvested thanks to the creation of solar cells that are transparent to visible light.

Windows typically transmit 55–90% of visible light. Richard Lunt and Vladimir Bulovic at the Massachusetts Institute of Technology in Cambridge have developed an organic solar cell made from chloroaluminium phthalocyanine and carbon-60, which absorb ultraviolet and near-infrared light. The researchers also added near-infrared mirror coatings to boost performance. The cell's efficiency of 1.7% is comparable to that of a similar opaque cell, 2.4%. The authors say that their device transmits enough visible light (55%) to be useful for architectural glass.

Appl. Phys. Lett. 98, **113305 (2011)**

NATURE.COM

For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

POLICY

Smallpox stocks

The World Health Organization (WHO) has failed to decide when to destroy the world's last two remaining stocks of the virus that causes smallpox. A meeting of the WHO's decision-making body, the World Health Assembly, this week was supposed to produce a deadline, but the organization ended up deferring judgement until the 67th assembly meeting, in 2014. A US resolution calling for the stocks to be maintained for at least another five years ran into opposition led by Iran. See go.nature.com/7gmgck for more.

Three Gorges Dam

In an unusually frank assessment of problems caused by the controversial Three Gorges Dam in central China, the State Council, China's cabinet, has pledged to curb environmental deterioration in the area, and to tackle the pollution of water supplies downstream in the Yangtze River. Other problems that the council plans to address include the dam's potential to cause seismic disasters, and its effects on biological diversity. The 18 May statement came during a severe drought in provinces along the middle and lower reaches of the Yangtze that has devastated farmland and left millions of people short of drinking water. See go.nature.com/auffmx for more.

Carbon targets

On 17 May, Britain extended existing pledges to limit greenhouse-gas emissions beyond 2020. It plans to cut emissions to 50% below 1990 levels between 2023 and 2027. The establishment of such targets is required under 2008 legislation that mandates setting 'carbon budgets' for consecutive five-year periods to 2050 — by which time levels should have been cut by 80%. And in Australia, an independent climate advisory body established by the government in February has published its first report. The 23 May study from the Climate Commission, *The Critical Decade,* urges immediate action to cut carbon emissions.

J. STANMEYER/VII/CORBIS

Deforestation surges in the Amazon

Brazil's environment minister Izabella Teixeira has vowed to crack down harder on loggers clearing trees in the Amazon rainforest, after a sudden rise in deforestation. On 18 May, Brazil's National Institute for Space Research (INPE) released satellite data recording that 593 square kilometres of forest had been cleared in March and April, a 473% increase over the 103.5 square kilometres cut down in the same period last year. Much of the clearing occurred in the state of Mato Grosso (pictured, in 2008), where soya-bean farming is common. At a crisis meeting, Teixeira said it was too early to tell whether the surge related to anticipated changes in legislation governing forest preservation, as environmentalists claim.

HIV scandal

The last plaintiff suing Japan's government and five biomedical companies over HIV infection caused by tainted blood products settled last week for ¥28 million (US$340,000) in damages. Since 1989, nearly 1,400 patients — mostly haemophiliacs — have sued after being infected in the 1980s by blood coagulants that were not treated to kill viruses. In 1996, Naoto Kan, then health minister and now prime minister of Japan, admitted partial government responsibility in the scandal (see *Nature* **379,** 663; 1996). The court cases may now be closed, but the patients are stuck with the virus. The health ministry says that it will continue to support their treatment.

BUSINESS

Hepatitis approvals

As expected, the US Food and Drug Administration has approved what is only the second drug to directly target the hepatitis C virus. Telaprevir (Incivek), marketed by Vertex Pharmaceuticals in Cambridge, Massachusetts, was given the green light on 23 May — 10 days after the agency approved boceprevir (Victrelis) made by Merck of Whitehouse Station, New Jersey.

TEPCO's losses

The operators of Japan's stricken Fukushima nuclear power plant announced a net

loss of ¥1.25 trillion (US$15.3 billion) for the year ending 31 March, because of expenses set aside to deal with nuclear clean-up. The president of the Tokyo Electric Power Company (TEPCO), Masataka Shimizu, resigned after the figures were released on 20 May. He has been replaced by managing director Toshio Nishizawa. The company's share price has dropped by more than 80% since the earthquake and tsunami on 11 March.

Takeda drug deal

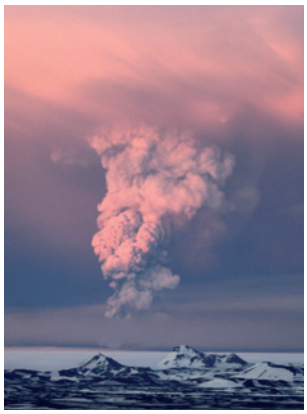
Japanese drug giant Takeda will buy Swiss drug maker Nycomed for €9.6 billion (US$13.6 billion), the firms announced on 19 May. Takeda's chief executive Yasuchika Hasegawa said Nycomed, based in Zurich, would give the Osaka-based company a better presence in Europe and emerging markets. Its last major deal was an $8.8-billion acquisition of biotech firm Millennium Pharmaceuticals, based in Cambridge, Massachusetts, in 2008.

EVENTS

Iceland's volcano

A year after the eruption of Eyjafjallajökull sent ash plumes across Europe, closing down airspace, another ice-covered volcano began

J. GUSTAFSSON/AP



erupting on 21 May (pictured). Grímsvötn spewed a plume of material some 20 kilometres into the sky, making the event far more powerful than the eruption of Eyjafjallajökull (which reached about 8 kilometres), but — thanks largely to prevailing winds — it is expected that there won't be such a large impact on European flights. Grímsvötn is Iceland's most active volcano, last erupting in 2004.

RESEARCH

Einstein telescope

A network of European researchers released designs on 19 May for an ultra-sensitive gravitational-wave observatory. The 'Einstein telescope', to be constructed around 2025, would be ten times more sensitive than even second-generation detectors expected to come online around 2015, such as the US Advanced LIGO experiment in Hanford, Washington State, and Livingston, Louisiana. It could also study in detail the interiors of sources producing gravitational waves. So far, no detectors have directly spotted gravitational waves — ripples in space-time thought to be produced by dramatic events such as the merger of black holes or neutron stars. See go.nature.com/apeqr5 for more.

Genome grants

The US National Institutes of Health's ENCODE (Encyclopedia of DNA Elements) programme, which aims to catalogue all the functional elements of the human genome, has been granted a US$123-million expansion. On 16 May, an advisory council to the National Human Genome Research Institute, which runs the programme, approved the money to allow the agency to develop calls for research proposals. Elise Feingold, director of ENCODE, told the council that completing the picture of RNA function was of particular interest.

PEOPLE

Nobel campaign

A group of 18 Nobel laureates have put their weight behind a campaign for sustainable development. Concluding the third Nobel Laureate Symposium on Global Sustainability on 16–19 May in Stockholm, the laureates signed a memorandum stating that economic and social development should go hand in hand with environmental protection. The document also calls for a major research initiative to better understand global sustainability. It was handed to the United Nations, which is preparing for a key conference on sustainable development in Rio de Janeiro, Brazil, in June 2012. See go.nature.com/9cmhbv for more.

Royal Society intake

Among 44 fellows elected to the Royal Society in London on 20 May were Nobel-prizewinning graphene researcher Kostya Novoselov of the University of Manchester; Mark Walport, head of the Wellcome Trust; and Bob Watson, a former chair of the Intergovernmental Panel on Climate Change. See go.nature.com/bmauut for more.

COMING UP

25–27 MAY
A host of eminent researchers speak at a free-to-attend conference on 'Transforming the future of energy', hosted by the US Department of Energy in Washington DC.
go.nature.com/zjrmew

29 MAY–2 JUNE
A world congress devoted to understanding the biological pathology behind psychiatric disorders is held in Prague.
go.nature.com/uxjkz1

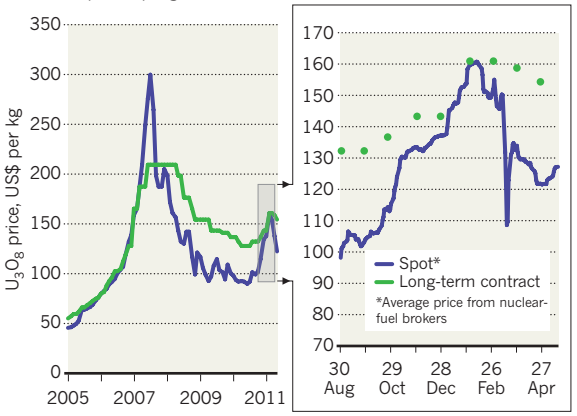
TREND WATCH

The price of uranium oxide — the raw material for uranium fuel — climbed steadily last year after an earlier price collapse. But since the disaster at the Fukushima Daiichi nuclear plant in Japan, the spot price has dropped by about 16%, reflecting uncertainty about prospects for nuclear energy. Ux Consulting, headquartered in Roswell, Georgia, has taken 10% off its projection of cumulative demand for uranium oxide to 2030, which now stands at around 2.3 billion kilograms.

SOURCE: UX CONSULTING

URANIUM STUTTERS AFTER FUKUSHIMA

The price of uranium oxide (U_3O_8) has plunged on fears of a slowdown in nuclear power programmes.



NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS

RUSSIA Resurgence of funding helps to lure expatriates home **p.428**

GENETICS Excitement and doubts over RNA editing claims **p.432**

VACCINES The high-stakes search for rare but real side effects **p.436**

VACCINES Revitalizing the quest for an HIV vaccine **p.439**

UNICEF/NYHQ2011-0198/ZAIDI

Children are still being disabled by polio in Pakistan despite years of effort to eradicate the disease.

PUBLIC HEALTH

Polio clings on in Pakistan

Fears grow that health-service reforms may let virus flourish, just as the global eradication effort reaches its endgame.

BY EWEN CALLAWAY

Asif Ali Zardari, the president of Pakistan, held a meeting last month to tackle yet another setback for his troubled nation. The focus was neither Al-Qaeda's resurgence nor poor diplomatic relations with the United States — it was polio, a crippling disease that has been wiped out in almost every other part of the world.

As an assessment of the global polio eradication campaign by its Independent Monitoring Board warns, "Pakistan risks being the country that prevents global polio eradication" (see go.nature.com/7tnvcg). After a spike in cases in Pakistan last year, a new emergency action plan to tackle the threat is off to a slow start with at least 36 confirmed cases this year. Meanwhile, constitutional reforms that will eliminate the country's central health ministry could slow efforts to turn the tide.

The push to make polio the second human pathogen after smallpox to be eradicated began in 1988. That year, an estimated 350,000 people developed poliomyelitis, an insidious infection that attacks the nervous system and can render patients paralysed within hours. The Global Polio Eradication Initiative, a public–private partnership led by the World Health Organization (WHO), hoped to finish the campaign by 2000. Yet many countries in Africa and central Asia did not begin eradication until the mid-1990s. Pakistan, meanwhile, successfully reduced its burden from 1,155 cases in 1997 to 28 in 2005.

By the mid-2000s, fewer than 2,000 people worldwide contracted polio each year, with the vast majority of cases occurring in Nigeria and India, where the campaign faced obstacles including vaccine boycotts. Public-awareness and vaccination campaigns have beaten back the disease since then: India has recorded just one case so far this year, and Nigeria eight.

"India and Nigeria decided to put on the retro-boosters and they dropped their cases from the thousands to below the levels of Pakistan," says Bruce Aylward, an assistant director-general at the WHO and head of the polio eradication campaign. "All of a sudden Pakistan is in the glare."

Pakistan is one of four countries in which polio is still endemic (see 'Stubborn holdouts'), but with 144 confirmed cases in 2010, it is the only country in which polio is making a comeback. The worsening situation in Pakistan could put gains elsewhere at risk. "As long as there's polio in any one country it's a threat to every country in the world," says David Heymann, chairman of the board of the UK Health Protection Agency and a former assistant director-general of the WHO involved in polio eradication.

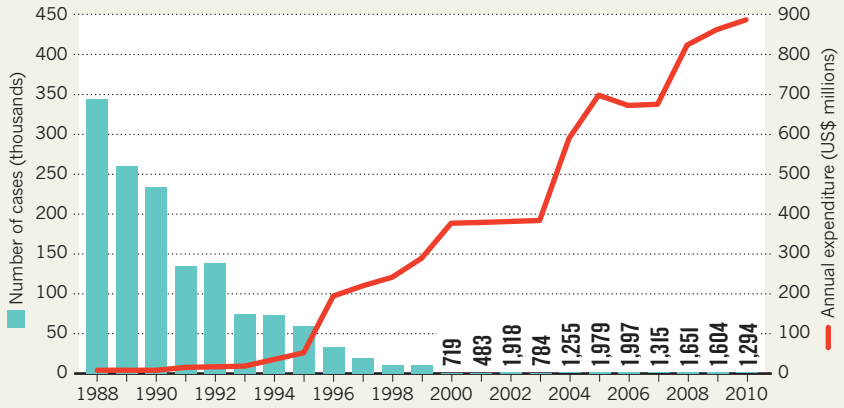
Climate, population density and other factors make the virus particularly infectious in Pakistan, Aylward says, requiring child vaccination rates of more than 95% to prevent its rapid spread. Yet immunization rates and surveillance are weak in Pakistan, according to the report released last month by the Independent Monitoring Board. Last year's floods displaced millions of people, spreading the virus and undercutting vaccination efforts. Poor security and religious opposition in the country's Federally Administered Tribal Areas may partly explain low vaccination rates there, but not the polio hotspot in the southern city of Karachi. Unaccountable local government officials and poor health infrastructure also plague many districts where polio has held on, according to the monitoring board report and WHO officials.

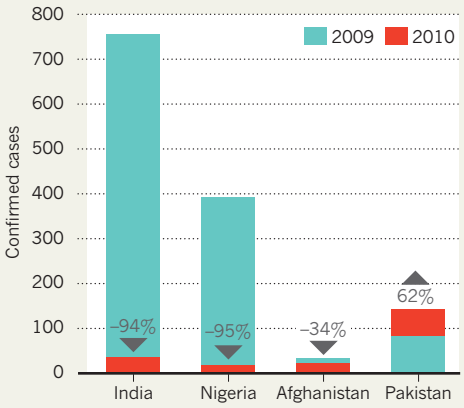
Donald Henderson of the Center for Biosecurity at the University of Pittsburgh Medical Center in Baltimore, Maryland, who led the ▶

VACCINES
New promise, old doubts
nature.com/vaccines

STUBBORN HOLDOUTS

Progress on eradicating polio has plateaued since 2000, with Pakistan proving the toughest nut to crack.





SOURCE: IMB REPT./GLOBAL POLIO ERADICATION INITIATIVE

WHO's war on smallpox in the 1960s and 1970s, says that smallpox also lingered in Pakistan long after it was eradicated from India and Bangladesh, which had similar natural obstacles to eradication. "To be experiencing polio is kind of a paradox," he says. "There's something seriously wrong with their health services."

Pakistan's government has not ignored the problem. In January, the country rolled out an emergency action plan, with the goal of halting virus transmission by the end of 2011. The plan seeks greater accountability at all levels of government to boost immunization rates, and it calls for health officials to work with the military and local leaders in tribal areas to build support for vaccination. Frustrated with the lack of progress reported at the April meeting, President Zardari also ordered an investigation into recent polio outbreaks in Sindh and Balochistan, and formed a new oversight committee to keep him personally informed about eradication efforts.

Yet some fear that looming constitutional reforms could make it harder for Pakistan to exterminate polio. The country is set to devolve its Ministry of Health by the end of June, part of a long-delayed move to transfer more power to provincial governments. Bill Gates, whose Bill & Melinda Gates Foundation in Seattle, Washington, funds polio eradication programmes, has expressed concern about the changes directly to Zardari. Sania Nishtar, a health-policy expert who heads Heartfile, an independent health-policy think tank in Islamabad, worries that without a central health authority to coordinate international donors' efforts and vaccine procurement and distribution, Pakistan's polio campaign will suffer.

Aylward, however, believes that the constitutional change is a "manageable risk". With eradication funding in Pakistan expected to reach US$137.5 million over the next two years, he wants to see clear signs of success. He hopes that the campaign will stop transmission of the virus by the end of 2012. In the meantime, after a couple of large vaccination campaigns, officials will assess whether the virus has vanished from some areas of Pakistan, and will look for a decline in the genetic diversity of the remaining virus — that would suggest that polio is on its last legs.

Simply blaming Pakistan for inaction is unhelpful, Aylward adds. "Everybody wants a whipping boy," he says. "Let's help them get this finished. This is our Alamo." ■ SEE COMMENT P.446

FUNDING

Russia revitalizes science

Researchers drawn by 'mega-grants' find rewards and frustrations in equal measure.

BY QUIRIN SCHIERMEIER

Siberia, of all places? Ernst-Detlef Schulze's wife rolled her eyes when her husband agreed to lead a major ecosystem study in the Yenisey region in the heart of Russia's eastern vastness. At first, Schulze, the founding director of the prestigious Max Planck Institute for Biogeochemistry in Jena, Germany, had been hesitant himself — but sensing a unique opportunity to study how the Arctic tundra and boreal forests store and release carbon, he decided to pack his bags.

The German carbon-cycle expert is one of 40 foreign or expatriate Russian scientists working in the West who last year received a new type of grant to bring their expertise to Russian universities. The 12-billion-rouble (US$428-million) 'mega-grant' programme is part of Russia's attempt to strengthen research at its neglected universities and modernize the country's science and economy at large (see *Nature* **465,** 858; 2010).

The ambitious plan is clear evidence that research money is now flowing generously in Russia, where a once-vast scientific workforce shrank dramatically in the years after the collapse of the Soviet Union (see 'After the fall'). The will to revitalize science is real, says Schulze. Alas, Kafkaesque bureaucracy and a thicket of often-opaque regulations have survived the changes.

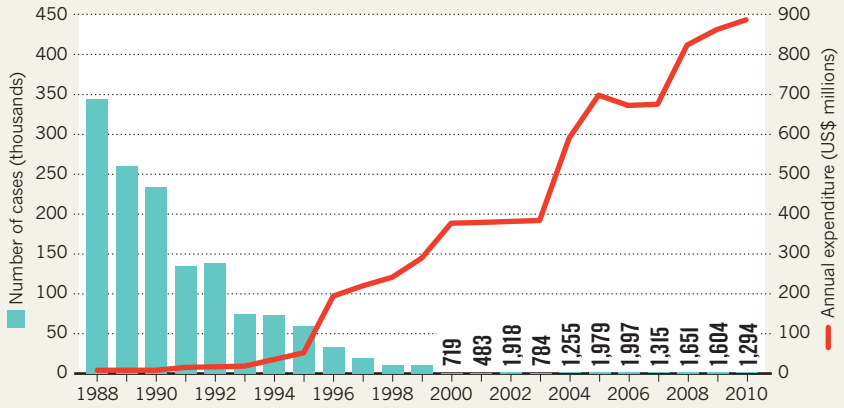
The Russian government is trying to smooth the way. This week, following a letter of complaint from scientists at Moscow State University (MSU), President Dmitry Medvedev met grant recipients to discuss the problems they have experienced, and promised to address them. But doing science in Russia remains a challenging, and often frustrating, mission for anyone unfamiliar with the country's idiosyncrasies, says Schulze.

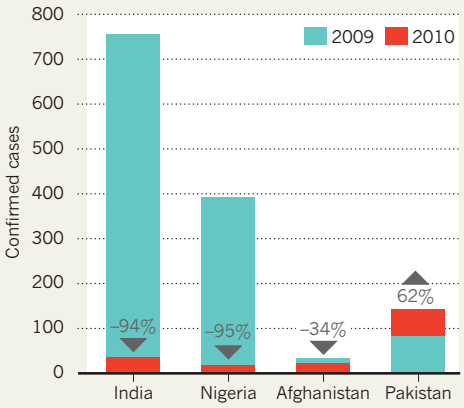
"The administrative, legal and academic environment can be perplexing," he says. "Essentially, you need very good contacts and a great deal of local support. If you arrive unprepared — thinking only that it would be cool to do some science in Siberia — you're lost."

Fortunately, Schulze does enjoy plenty of support. During his long career he has published more than 20 papers with

STUBBORN HOLDOUTS

Progress on eradicating polio has plateaued since 2000, with Pakistan proving the toughest nut to crack.





SOURCE: IMB REPT./GLOBAL POLIO ERADICATION INITIATIVE

WHO's war on smallpox in the 1960s and 1970s, says that smallpox also lingered in Pakistan long after it was eradicated from India and Bangladesh, which had similar natural obstacles to eradication. "To be experiencing polio is kind of a paradox," he says. "There's something seriously wrong with their health services."

Pakistan's government has not ignored the problem. In January, the country rolled out an emergency action plan, with the goal of halting virus transmission by the end of 2011. The plan seeks greater accountability at all levels of government to boost immunization rates, and it calls for health officials to work with the military and local leaders in tribal areas to build support for vaccination. Frustrated with the lack of progress reported at the April meeting, President Zardari also ordered an investigation into recent polio outbreaks in Sindh and Balochistan, and formed a new oversight committee to keep him personally informed about eradication efforts.

Yet some fear that looming constitutional reforms could make it harder for Pakistan to exterminate polio. The country is set to devolve its Ministry of Health by the end of June, part of a long-delayed move to transfer more power to provincial governments. Bill Gates, whose Bill & Melinda Gates Foundation in Seattle, Washington, funds polio eradication programmes, has expressed concern about the changes directly to Zardari. Sania Nishtar, a health-policy expert who heads Heartfile, an independent health-policy think tank in Islamabad, worries that without a central health authority to coordinate international donors' efforts and vaccine procurement and distribution, Pakistan's polio campaign will suffer.

Aylward, however, believes that the constitutional change is a "manageable risk". With eradication funding in Pakistan expected to reach US$137.5 million over the next two years, he wants to see clear signs of success. He hopes that the campaign will stop transmission of the virus by the end of 2012. In the meantime, after a couple of large vaccination campaigns, officials will assess whether the virus has vanished from some areas of Pakistan, and will look for a decline in the genetic diversity of the remaining virus — that would suggest that polio is on its last legs.

Simply blaming Pakistan for inaction is unhelpful, Aylward adds. "Everybody wants a whipping boy," he says. "Let's help them get this finished. This is our Alamo." ■ SEE COMMENT P.446

FUNDING

Russia revitalizes science

Researchers drawn by 'mega-grants' find rewards and frustrations in equal measure.

BY QUIRIN SCHIERMEIER

Siberia, of all places? Ernst-Detlef Schulze's wife rolled her eyes when her husband agreed to lead a major ecosystem study in the Yenisey region in the heart of Russia's eastern vastness. At first, Schulze, the founding director of the prestigious Max Planck Institute for Biogeochemistry in Jena, Germany, had been hesitant himself — but sensing a unique opportunity to study how the Arctic tundra and boreal forests store and release carbon, he decided to pack his bags.

The German carbon-cycle expert is one of 40 foreign or expatriate Russian scientists working in the West who last year received a new type of grant to bring their expertise to Russian universities. The 12-billion-rouble (US$428-million) 'mega-grant' programme is part of Russia's attempt to strengthen research at its neglected universities and modernize the country's science and economy at large (see *Nature* **465,** 858; 2010).

The ambitious plan is clear evidence that research money is now flowing generously in Russia, where a once-vast scientific workforce shrank dramatically in the years after the collapse of the Soviet Union (see 'After the fall'). The will to revitalize science is real, says Schulze. Alas, Kafkaesque bureaucracy and a thicket of often-opaque regulations have survived the changes.

The Russian government is trying to smooth the way. This week, following a letter of complaint from scientists at Moscow State University (MSU), President Dmitry Medvedev met grant recipients to discuss the problems they have experienced, and promised to address them. But doing science in Russia remains a challenging, and often frustrating, mission for anyone unfamiliar with the country's idiosyncrasies, says Schulze.

"The administrative, legal and academic environment can be perplexing," he says. "Essentially, you need very good contacts and a great deal of local support. If you arrive unprepared — thinking only that it would be cool to do some science in Siberia — you're lost."

Fortunately, Schulze does enjoy plenty of support. During his long career he has published more than 20 papers with

Russian colleagues, and some of his long-term collaborators are now helping him to get settled at his host institute, the Siberian Federal University in Krasnoyarsk.

"Our faculty and our students are very lucky to have Detlef doing research here," says Evgeny Vaganov, the rector of the university. "We do everything we can to support him with the admittedly excessive paperwork."

Vaganov has commissioned a professor of economics to support Schulze full-time with the logistics of his planned fieldwork, with the procurement and import of equipment, and in his interactions with local and state authorities. Even so, the purchase of every single piece of equipment for Schulze's project had to be approved by the Russian security services, who Schulze says are usually suspicious of his efforts. "If it's new it's bad," he says. "That's the kind of mindset, unfortunately, that is hampering Russia's intellectual and technological renaissance. It is blocking curiosity and inventive genius."

Similar problems crop up when buying chemical reagents or transporting biological samples into or out of Russia. "In Sweden, if I need a chemical I order it, and I'll get it in a matter of days," says Boris Zhivotovsky, a cell biologist at the Karolinska Institute in Stockholm and a holder of one of the grants, which he is using to set up a lab and PhD programme at MSU. "In Russia, you need to apply for, and put out to tender, every little thing you need for your research," he says. "It takes three months or more until you get what you need."

Other grant recipients worry about the time limits they face. "My main concern is what will happen after the two-year funding period," says Stanislav Smirnov, a mathematician at the University of Geneva, Switzerland, and winner of the 2010 Fields Medal, who is using his grant to work at St Petersburg State University. "In mathematics, it can easily take two years to get one paper published. If this programme is to have a lasting impact, I think there would need to be a mechanism for expanding it beyond the short two-year funding period. Otherwise I fear the investment will just peter out."

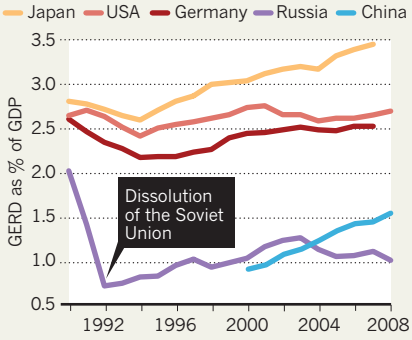
The mega-grant programme is part of a much wider drive to boost science and innovation. Four years ago, the Russian government launched an ambitious 318-billion-rouble nanotechnology initiative (see *Nature* **461,** 1036–1037; 2009). And a science and innovation centre currently being built in Skolkovo near Moscow will host Russian and foreign entrepreneurs and companies, including heavyweights such as Siemens and Bosch. Commercial research and development activities are expected to begin next year in five key areas — energy, information and communication, biotechnology, space and nuclear technologies.

G. S. DE BLONSKY/ALAMY

Scientists at Moscow State University have been helped by colleagues but hindered by bureaucracy.

AFTER THE FALL

Following the collapse of the Soviet Union, the Russian Federation has slowly ramped up its investment in research and development.



GERD, gross expenditure on R&D; GDP, gross domestic product

SOURCE: UNESCO/OECD

But the poorly developed technology-transfer process at Russia's academic institutions is proving to be a major problem, says grantholder Alexander Kabanov, a biochemist at the Nebraska Medical Center in Omaha, who is setting up a drug-discovery group at MSU. In the absence of a well established domestic patent system (and with no budget available at universities for filing patents in the European Union or the United States), collaborators in intellectual-property-intensive fields such as Kabanov's risk leaving their inventions inadequately protected.

Russia's regulations governing the welfare of animals used in experiments are also less stringent than in many other countries, which can cause problems for scientists who want to publish in international journals. "It's a touchy issue," says Kabanov. "It is important that what we do is not considered questionable in the West. When it comes to animal experiments, we need to apply the same protocols and ethical standards as we do everywhere else."

But he emphasizes that the mega-grant programme is already bearing fruit — and not only for Russia. Those returning to the country are "also benefiting as scientists and intellectuals" from the country's fine scholarly tradition and respect for science, he says, citing his new colleagues' willingness to share their knowledge.

Meanwhile, the Russian science ministry has already confirmed its commitment to the scheme by announcing a second call for proposals. "I expect they'll get at least twice as many applications as in the first round," says Vaganov, who is currently helping researchers to prepare six applications for projects to be hosted by the Siberian Federal University. "Despite some problems," he says, "this programme is a wonderful opportunity to uplift Russian science and get our students in touch with world-class research happening at our doorsteps." ■

MORE ONLINE

P. CREAN A / ALAMY

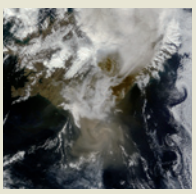
TOP STORY

Integrating renewables into the grid is not as hard as was thought go.nature.com/q2shdz

MORE NEWS

- Should flame-retardant chemicals be banned? go.nature.com/7winxx
- AIDS mortality drops in China go.nature.com/bng6d5
- Don't worry about the range of electric vehicles — many people don't drive that far go.nature.com/jnxkzb

ON THE BLOG



Here we go again. Is Grímsvötn anything like Eyjafjallajökull? go.nature.com/wge3je

NASA MODIS

REUTERS/D. BALIBOUSE

The World Health Assembly, meeting last week, described the WHO as archaic and overextended.

GLOBAL HEALTH FUNDING

Revamp for WHO

Critics call for restructuring of world health body, together with greater focus on fewer issues.

BY DECLAN BUTLER

The top decision-making body at the World Health Organization (WHO) — the World Health Assembly — last week backed reforms that might bring the biggest changes to the agency in its 63-year history.

Concerns about the WHO's performance stretch back decades, but the current harsh financial climate and an altered global-health landscape have brought the need for reform to a head. The WHO's member states now want the overextended agency to focus its activities on a far smaller number of core areas. But some experts think the reforms are unlikely to go far enough, and are calling for an overhaul of the agency's structure and governance.

Just a decade ago, the WHO was the only major global-health player, but it is now struggling to assert its *raison d'ĕtre* against a plethora of powerful and often better-funded new players, among them the Global Fund to Fight AIDS, Tuberculosis and Malaria, the GAVI Alliance and the Bill & Melinda Gates Foundation. Funding for global health has grown enormously over the past decade, from US$5.7 billion in 1990 to $26.9 billion last year, but this money has largely bypassed the WHO because of donors' lack of confidence in the agency (see 'Lost in the crowd'). Indeed, the WHO ran a $300-million deficit in 2010.

The agency suffered another blow last week when, in Geneva, Switzerland, the World Health Assembly's giant annual gathering of health ministers from the WHO's 193 member states fixed its budget for the next two years at $3.96 billion, almost $1 billion less than the $4.8 billion requested by the WHO. One immediate impact is to force the agency to shave some 300 of the 2,400 jobs at its Geneva headquarters.

Margaret Chan, the agency's director-general, last week attributed the WHO's financial woes to reductions in member states' contributions as a result of the global economic crisis. She also blamed the weakness of the dollar — the currency in which the WHO receives its contributions — against the Swiss franc, which it uses to pay staff and other costs at its headquarters. But member states see the crisis as an opportunity to modernize the agency, whose basic structures and procedures date from its creation in a very different world in 1948.

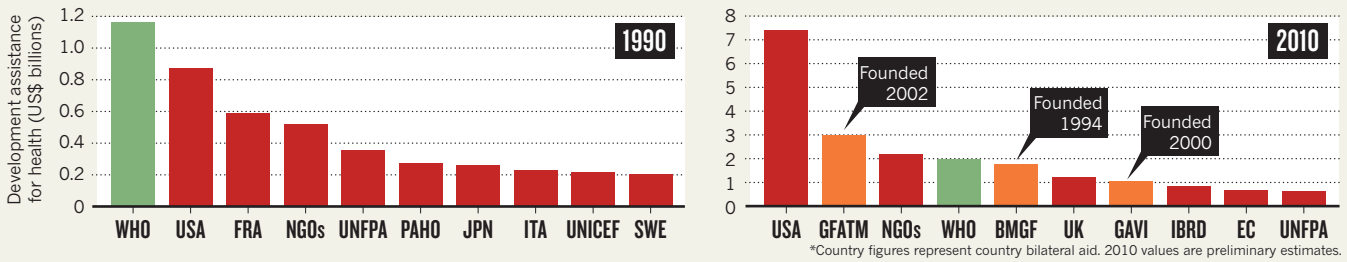
The reform proposal approved by the assembly last week lacks specifics, and is most significant as a green light for what will be complex negotiations. But its appraisal of the WHO is not glowing — it depicts the agency as archaic, overextended and lacking adequate assessment of its programmes and staff.

"The WHO has been trying to do too much; we need greater focus," concedes Andrew Cassels, director of Chan's office. He adds that the WHO under Chan welcomes the proposed reforms, which include calls for management change and for the agency to focus on data collection, developing standards for diseases, treatments and technologies, overseeing reponses to pandemics and longer-term health threats, and strengthening health systems. Cassels also embraces the proposal that the WHO flex its convening power — its capacity to bring together researchers, fieldworkers, public-health experts and politicians — to better coordinate the efforts of all players in global health.

Barry Bloom from the Harvard School of Public Health in Boston, Massachusetts, who authored a recent call for WHO reform in *Nature* (B. R. Bloom *Nature* **473,** 143–145; 2011), says that overall he is "very pleased" with the thrust of the proposals. Bloom applauds the priority given to the WHO's convening role, and a plan to open the agency up to independent external review for the first time. He and others also support a proposal to establish a

LOST IN THE CROWD

The World Health Organization once dominated global funding for health. Powerful new funding organizations and growing national aid programmes have surpassed it.



*Country figures represent country bilateral aid. 2010 values are preliminary estimates.

BMGF Bill & Melinda Gates Foundation; **EC** European Commission; **FRA** France; **GAVI** Alliance (formerly the Global Alliance for Vaccines and Immunisation); **GFATM** Global Fund to Fight AIDS, Tuberculosis and Malaria; **IBRD** International Bank for Reconstruction and Development; **ITA** Italy; **JPN** Japan; **NGOs** Other non-governmental organizations; **PAHO** Pan-American Health Organization; **SWE** Sweden; **UNFPA** United Nations Population Fund; **UNICEF** United Nations Children's Fund; **UK** United Kingdom; **USA** United States; **WHO** World Health Organization.

forum that would give non-governmental organizations, the drug industry and other stakeholders a say in shaping the WHO's policies. This is critical, say experts, as such stakeholders have key roles in global health. Member states fear a loss of their influence, however, and many observers are concerned at the risk of undue industry sway.

But some experts say the reform plans fail to explicitly address fundamental structural weaknesses in the WHO. One is its inability to set its own priorities. The agency has full control of only the 25% of its budget that comes from membership fees paid by member states. A whopping 75% comes from voluntary contributions — funds that donor countries often earmark for their own pet projects. This skews the WHO's priorities, says Lawrence Gostin, head of the WHO Collaborating Center on Public Health Law and Human Rights at Georgetown University in Washington DC. "Around 60% of the WHO's budget goes on infectious diseases, and just 3.9% on non-communicable diseases and 2.4% on injuries, yet these are a huge global health burden," he notes.

Cassels says that building greater trust in the WHO should result in fewer earmarks. But experts are also concerned that the proposals ignore an elephant in the room: the WHO's byzantine decentralized structure. Its six regional offices are autonomous to an extent seen in no other UN agency, electing their own heads, controlling large amounts of the WHO's funds and largely fixing their own policies. "At present, there are seven WHOs, not one," says Bloom.

"It's a birth defect" rooted in the WHO's creation from pre-existing regional bodies, says Kelley Lee, an expert on the WHO at the London School of Hygiene and Tropical Medicine. "It adds to the lack of cohesion and expansive agenda that spreads the WHO's limited resources ever more thinly." Cassels argues, however, that much improvement is possible within the existing decentralized organization. He says that Chan intends to take charge of an effort to "require much stronger monitoring of performance across all of the major offices".

Yet if deeper structural issues are not tackled, reforms will bring only "marginal" change, predicts Derek Yach, senior vice-president of global health and agriculture policy at PepsiCo, headquartered in New York State, who was a cabinet secretary to former WHO director-general Gro Harlem Brundtland. And Lee thinks that fixing the WHO won't be enough to bring coherence to global health funding. "I believe that the biggest problem is that there are now so many institutional players in global health, and that there is a need to reform the whole chaotic landscape, not just one organization. It's like fixing a flat tyre when the whole car needs to be tuned up." ■

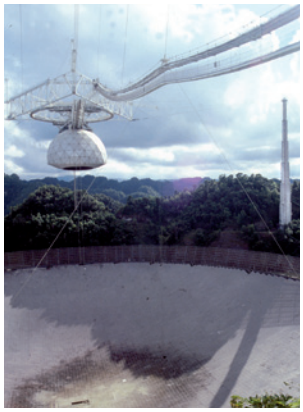
ASTRONOMY

Change rattles the world's biggest dish

After nearly half a century, Cornell University loses stewardship of the renowned Arecibo radio telescope.

BY EUGENIE SAMUEL REICH

As Earth's biggest 'ear' on the Universe, the giant 305-metre radio dish at Arecibo, Puerto Rico, has played a part in groundbreaking discoveries, searches for alien civilizations and the occasional Hollywood movie. Now a different sort of drama is shaking up the facility, with the news that Cornell University, which has managed Arecibo since the observatory was switched on 1963, has lost its bid to continue to do so. Instead, the US National Science Foundation (NSF) has offered the job — and the US$41.2-million five-year contract that goes with it — to a consortium that includes SRI International, a non-profit research institute based in Menlo Park, California; the Universities Space Research Association in Washington DC; and the Metropolitan University in Puerto Rico.



The Arecibo Observatory.

J.A. MANCHESTER/CUSTOM MEDICAL STOCK PHOTO/NEWSCOM

The decision means an abrupt switch in the status of about 100 scientists, engineers and support staff at the observatory, who will no longer be on the Cornell payroll. "All of our staff are somewhat disoriented because they've been here 20, 30, 40 years and never considered they wouldn't be Cornell employees," says Sixto González, assistant director of space and atmospheric science at Arecibo and the site's former director.

Rumours of the change first appeared on the Internet last week and have since been confirmed by Cornell. "We wish SRI good luck in living up to the incredibly high standards we have set," says Ira Wasserman, chairman of Cornell's astronomy department.

Robert Kerr, a former director of Arecibo who was principal investigator on the successful bid, says that, to ensure a smooth transition, for the first year nothing will change about how applications for time on the facility are processed. Beyond that, Kerr says, partners in the consortium are "likely to bring forward new ideas and new science".

One thing that will change right away is the location of the observatory's director. Under Cornell, Arecibo's director was based remotely at the university's campus in Ithaca, New York. Under the consortium, the director — expected to be Kerr — will work on site.

Héctor Arce, an astronomer at Yale University in New Haven, Connecticut, who was born and raised in Puerto Rico, says it was widely felt that Cornell didn't do enough to partner with other institutions on the island. "That might have hurt Cornell," he says. In addition to including the Metropolitan University on the bid, the consortium will sponsor faculty positions at the University of Puerto Rico and will create a commission that will study other ways in which local institutions can get more involved with Arecibo.

Donald Campbell, a Cornell astronomer and the current director of the observatory, says Cornell submitted a serious proposal based on partnerships with seven institutions, including the University of Puerto Rico.

But Juan Arratia, an electrical engineer at the Metropolitan University in San Juan, says that the Puerto Rican government supported the SRI consortium's bid through the government-owned Puerto Rico Industrial Development Company, the office of tourism and the Department of Education. Arratia, who led Metropolitan's part of the bid, expects that government sources will contribute around $5 million per year to the observatory. "That would be a real plus," says John Salzer, an astronomer at Indiana University Bloomington who works with data from Arecibo.

Astronomers rallied to support Arecibo after its closure was recommended in a 2006 NSF review. The facility is now likely to remain the most sensitive instrument of its kind for the foreseeable future. A 500-metre radio dish under construction in Guizhou, China, will not reach the higher frequencies that Arecibo can detect. The Square Kilometre Array, to be built in Australia or South Africa, would surpass Arecibo, but its construction is not expected to begin for several more years. ■

forum that would give non-governmental organizations, the drug industry and other stakeholders a say in shaping the WHO's policies. This is critical, say experts, as such stakeholders have key roles in global health. Member states fear a loss of their influence, however, and many observers are concerned at the risk of undue industry sway.

But some experts say the reform plans fail to explicitly address fundamental structural weaknesses in the WHO. One is its inability to set its own priorities. The agency has full control of only the 25% of its budget that comes from membership fees paid by member states. A whopping 75% comes from voluntary contributions — funds that donor countries often earmark for their own pet projects. This skews the WHO's priorities, says Lawrence Gostin, head of the WHO Collaborating Center on Public Health Law and Human Rights at Georgetown University in Washington DC. "Around 60% of the WHO's budget goes on infectious diseases, and just 3.9% on non-communicable diseases and 2.4% on injuries, yet these are a huge global health burden," he notes.

Cassels says that building greater trust in the WHO should result in fewer earmarks. But experts are also concerned that the proposals ignore an elephant in the room: the WHO's byzantine decentralized structure. Its six regional offices are autonomous to an extent seen in no other UN agency, electing their own heads, controlling large amounts of the WHO's funds and largely fixing their own policies. "At present, there are seven WHOs, not one," says Bloom.

"It's a birth defect" rooted in the WHO's creation from pre-existing regional bodies, says Kelley Lee, an expert on the WHO at the London School of Hygiene and Tropical Medicine. "It adds to the lack of cohesion and expansive agenda that spreads the WHO's limited resources ever more thinly." Cassels argues, however, that much improvement is possible within the existing decentralized organization. He says that Chan intends to take charge of an effort to "require much stronger monitoring of performance across all of the major offices".

Yet if deeper structural issues are not tackled, reforms will bring only "marginal" change, predicts Derek Yach, senior vice-president of global health and agriculture policy at PepsiCo, headquartered in New York State, who was a cabinet secretary to former WHO director-general Gro Harlem Brundtland. And Lee thinks that fixing the WHO won't be enough to bring coherence to global health funding. "I believe that the biggest problem is that there are now so many institutional players in global health, and that there is a need to reform the whole chaotic landscape, not just one organization. It's like fixing a flat tyre when the whole car needs to be tuned up." ■

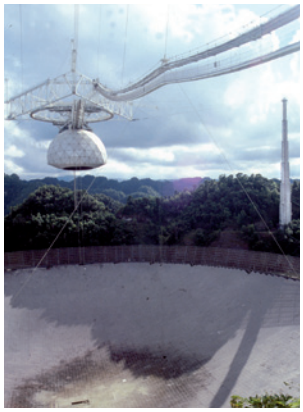
ASTRONOMY

Change rattles the world's biggest dish

After nearly half a century, Cornell University loses stewardship of the renowned Arecibo radio telescope.

BY EUGENIE SAMUEL REICH

As Earth's biggest 'ear' on the Universe, the giant 305-metre radio dish at Arecibo, Puerto Rico, has played a part in groundbreaking discoveries, searches for alien civilizations and the occasional Hollywood movie. Now a different sort of drama is shaking up the facility, with the news that Cornell University, which has managed Arecibo since the observatory was switched on 1963, has lost its bid to continue to do so. Instead, the US National Science Foundation (NSF) has offered the job — and the US$41.2-million five-year contract that goes with it — to a consortium that includes SRI International, a non-profit research institute based in Menlo Park, California; the Universities Space Research Association in Washington DC; and the Metropolitan University in Puerto Rico.



The Arecibo Observatory.

J.A. MANCHESTER/CUSTOM MEDICAL STOCK PHOTO/NEWSCOM

The decision means an abrupt switch in the status of about 100 scientists, engineers and support staff at the observatory, who will no longer be on the Cornell payroll. "All of our staff are somewhat disoriented because they've been here 20, 30, 40 years and never considered they wouldn't be Cornell employees," says Sixto González, assistant director of space and atmospheric science at Arecibo and the site's former director.

Rumours of the change first appeared on the Internet last week and have since been confirmed by Cornell. "We wish SRI good luck in living up to the incredibly high standards we have set," says Ira Wasserman, chairman of Cornell's astronomy department.

Robert Kerr, a former director of Arecibo who was principal investigator on the successful bid, says that, to ensure a smooth transition, for the first year nothing will change about how applications for time on the facility are processed. Beyond that, Kerr says, partners in the consortium are "likely to bring forward new ideas and new science".

One thing that will change right away is the location of the observatory's director. Under Cornell, Arecibo's director was based remotely at the university's campus in Ithaca, New York. Under the consortium, the director — expected to be Kerr — will work on site.

Héctor Arce, an astronomer at Yale University in New Haven, Connecticut, who was born and raised in Puerto Rico, says it was widely felt that Cornell didn't do enough to partner with other institutions on the island. "That might have hurt Cornell," he says. In addition to including the Metropolitan University on the bid, the consortium will sponsor faculty positions at the University of Puerto Rico and will create a commission that will study other ways in which local institutions can get more involved with Arecibo.

Donald Campbell, a Cornell astronomer and the current director of the observatory, says Cornell submitted a serious proposal based on partnerships with seven institutions, including the University of Puerto Rico.

But Juan Arratia, an electrical engineer at the Metropolitan University in San Juan, says that the Puerto Rican government supported the SRI consortium's bid through the government-owned Puerto Rico Industrial Development Company, the office of tourism and the Department of Education. Arratia, who led Metropolitan's part of the bid, expects that government sources will contribute around $5 million per year to the observatory. "That would be a real plus," says John Salzer, an astronomer at Indiana University Bloomington who works with data from Arecibo.

Astronomers rallied to support Arecibo after its closure was recommended in a 2006 NSF review. The facility is now likely to remain the most sensitive instrument of its kind for the foreseeable future. A 500-metre radio dish under construction in Guizhou, China, will not reach the higher frequencies that Arecibo can detect. The Square Kilometre Array, to be built in Australia or South Africa, would surpass Arecibo, but its construction is not expected to begin for several more years. ■

GENETICS

Evidence of altered RNA stirs debate

Sceptics question find that upends biology's 'central dogma'.

BY ERIKA CHECK HAYDEN

A funny thing happened on the way to the ribosome. That's the essence of a controversial paper concluding that messenger RNA — the molecular middleman that carries information from a cell's DNA to its protein-making machinery — is routinely and systematically altered by unknown mechanisms before its genetic instructions can be read. The paper, published in *Science* last week (M. Li *et al. Science* doi:10.1126/science.1207018; 2011), is already drawing pointed reviews from computational biologists, who cite possible flaws that could undermine the authors' claims.

If verified, the findings would require a rewrite of the 'central dogma' of molecular biology, which posits that the RNA transcripts that carry genetic information to the ribosome, where they are used as templates for protein assembly, are generally faithful matches to the original DNA. A revised version of the picture would include an 'RNA editing' step along the way, which replaces individual letters in the genetic code and changes the resulting proteins (see 'Unmatched messages'). Such a step would allow cells to generate much more diversity from the standard DNA tool kit than previously thought.

Vivian Cheung of the University of Pennsylvania in Philadelphia led the work, which involved examining the RNA transcripts and DNA sequences of 27 people who were sequenced in the 1000 Genomes Project and the International HapMap Project. The team found more than 10,000 sites in exons — regions of messenger RNA that have been transcribed from DNA — in which the DNA and RNA sequences did not match. The same mismatches occurred in different people, suggesting that they were not random mistakes in transcription. Cheung's team also found proteins made from the 'mismatched' RNAs.

"The big challenge now is to sort out the molecular mechanism."

"Once we saw that these differences were translated into protein sequences, we were pretty certain that they were biologically derived," Cheung says.

RNA editing — a process that changes the identity of an RNA base after it has been transcribed from a DNA sequence — is not a new discovery. An enzyme called ADAR, for instance, induces mismatches in human cells by replacing the base adenosine with another molecule that is then read as guanine when the RNA is used to code for a protein. RNA editing also occurs in plants and human parasites.

But the extent of RNA editing posited by the *Science* paper is extraordinary; its authors estimate that each person has about 1,065 mismatches — sites the authors call "RNA–DNA differences", or RDDs. Some of the mismatches involve base changes that are not produced by known RNA-editing mechanisms, suggesting that undiscovered mechanisms are at work.

"This suggests a completely different layer of gene regulation at the RNA level," says molecular biologist Kazuko Nishikura at the Wistar Institute in Philadelphia. "The big challenge now is to sort out the molecular mechanism for how these RNA sequence alterations can be achieved."

Others remain sceptical. Comparative genomicist Lior Pachter at the University of California, Berkeley, has studied how the high-throughput sequencing machines that Cheung's team used to sequence RNA make systematic errors when sequencing DNA and RNA. He says that some of Cheung's mismatches occur at sites that are prone to systematic RNA sequencing errors, but others do not.

And in a post on the blog 'genomes unzipped' on 20 May, Joe Pickrell, a graduate student working with human geneticist Jonathan Pritchard at the University of Chicago, Illinois, described another potential source of error. Pickrell said that multiple regions of similar DNA in the human genome can make it difficult to trace the origin of a short stretch of RNA to a specific DNA sequence, creating the illusion of DNA–RNA differences. "If the authors are accidentally attributing RNA from two different regions of the genome to the same DNA region, they could falsely infer RNA editing," Pickrell said. "I think many of their results could be the result of errors in identifying the correct genomic origin of their sequencing reads."

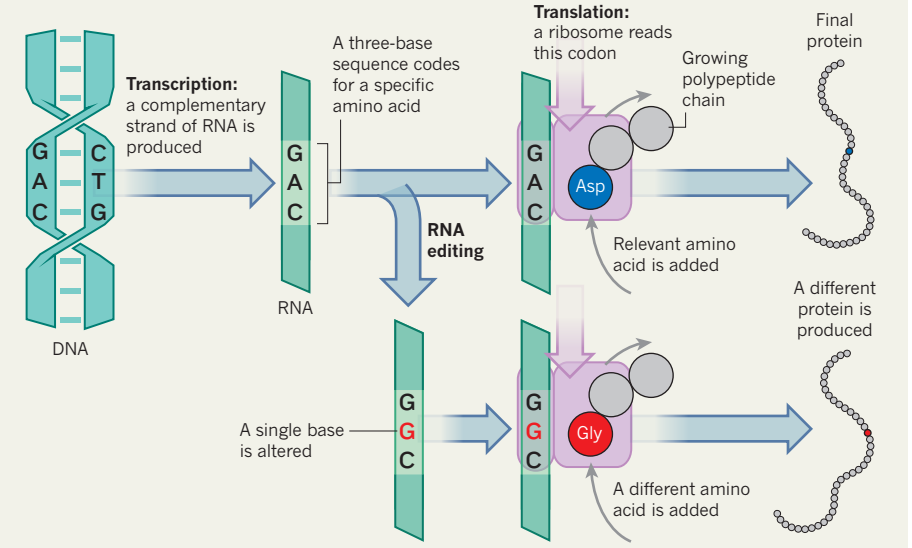
Other researchers are combing through their own data and waiting to see the results of follow-up work that will determine whether the concerns raised by Pachter, Pickrell and others are valid. Meanwhile, Chueng says, "we are glad to see that our colleagues are already using our data".

If confirmed, Cheung's work has important implications for biology and for the way that researchers study genomics. Chris Gunter, director of research affairs at the Hudson-Alpha Institute for Biotechnology in Huntsville, Alabama, says that RNA editing might have implications for the genetic origins of disease, if it turns out that the control of how much editing occurs is inherited.

"This could make our jobs as geneticists more problematic and more interesting," she says. ■

UNMATCHED MESSAGES

The swapping out of a single base in a section of messenger RNA alters the genetic information transcribed from the DNA before it is translated into a protein.



NATURE.COM
For more on the complexity of the genome, see:
go.nature.com/unyfgv

TRANSLATIONAL RESEARCH

Therapeutic success stifles medical progress

Drug development loses momentum as patients shun clinical trials for tried and tested treatments. Could payment for participation be the answer?

BY HEIDI LEDFORD

Is medical research a victim of its own success? A surprising economic analysis suggests that each new medical innovation may make the next more difficult to achieve, because patients prefer to stick with proven — though potentially inferior — treatments rather than trying something new. Good effectively becomes the enemy of great.

The finding confirms the experience of many medical researchers struggling to recruit patients for their next clinical trial. But the solution proposed by the study's authors — to pay trial participants higher stipends — makes some clinicians queasy.

The analysis[1], published this month by the National Bureau of Economic Research based in Cambridge, Massachusetts, shows that the percentage of HIV-infected men who enrolled in clinical trials plummeted immediately after a regimen of antiretroviral drugs known as HAART hit the market in 1996 (see 'The price of success'). The precipitous decline, seen in data from the Multicenter AIDS Cohort Study of around 7,000 homosexual and bisexual men in four US cities from 1984 to 2005, occurred even as federal funding for HIV research nearly doubled. HAART simply worked so well, says Anup Malani, a law professor at the University of Chicago, Illinois, and an author of the new study, that patients were no longer motivated to sign up for clinical trials as a way to gain access to better treatments.

Malani and his co-author Tomas Philipson, also of the University of Chicago, believe similar problems arise whenever a new drug significantly improves the treatment of a particular disease, leading to a decline in pharmaceutical productivity. "Innovations of today increase the cost of innovations tomorrow," Malani says. "And that means the cost of the next-generation drug is going to rise."

HAART represented a therapeutic revolution for a group of patients with few options, but Malani's economic models predict that incremental therapeutic advances would also affect clinical-trial enrolment for other diseases, although the size of the effect may not be as large. Richard Schilsky, an oncologist at the University of Chicago Medical Center, notes that a decade ago, there was only one treatment approved by the US Food and Drug Administration for kidney cancer. Now there are seven, and clinicians working on the disease struggle to find patients for new trials. "There are so many options that patients are not flocking to get into clinical trials like they used to," he says.

Failure to recruit sufficient participants is a common reason for stopping a clinical trial. The struggle to find enough people is also one reason that companies are increasingly performing clinical trials in developing nations where infrastructure and labour is cheaper, and patients with limited resources are more willing to sign on to a trial as a way to access expensive drugs.

In the United States, it is common to pay healthy volunteers who enrol in trials to test a drug's safety. But when it comes to testing drug efficacy in sick patients, payment is often limited to compensation for incidental expenses, such as travel costs or parking fees. In fact, studies[2] have suggested that these participants may be undercompensated. Investigators and ethics committees tend to underestimate the real cost to participants, which may also include child care and time away from work.

Yet some fear that higher compensation will subject sick patients to undue inducement, which is forbidden under US federal regulations governing research on humans. "There's a real moral dilemma," acknowledges Malani. "We allow people to become race car drivers and get a wage premium for that, but I think we have to be careful when somebody is sick and making that kind of a decision."

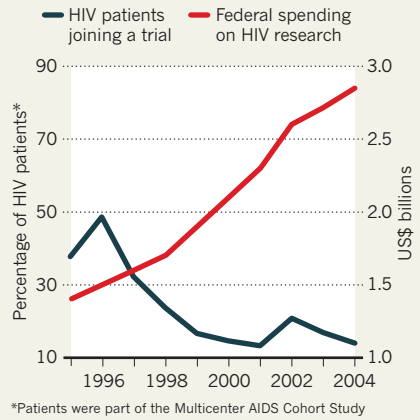
"I don't think the solution is necessarily paying patients," Schilsky says, despite the challenges of recruiting trial participants. But others feel that cash does not always equal coercion. "There is undue concern about undue inducement," says Robert Klitzman, a psychiatrist and bioethicist at Columbia University in New York. "I think it is overly feared that payment will be coercive."

Malani suggests that one way to ease these fears may be to adopt more protections, such as education programmes to ensure that people know the risks of participating in a clinical trial. Investigators could also be asked to certify that participants are not taking on too large a health risk for the monetary gain.

Refusing to pay patients a fair stipend is in some ways hypocritical, says Elizabeth Ripley, a nephrologist and senior chair of the ethics committees at Virginia Commonwealth University in Richmond. Clinicians conducting a trial stand to gain prestige and publications, and they often receive a payment for each patient enrolled, she notes. Institutions also get financial support for hosting the trial. Ultimately, the benefits of doing the study — which include benefits to society — have to outweigh the effort and the time. "We as researchers don't do studies if we don't get anything out of it," she says.

"Refusing to pay patients a fair stipend is in some ways hypocritical."

Malani says that he and Philipson realized they were stepping into an ethical minefield when they broached the issue. But delaying drug discovery is ethically troubling as well, Malani says. "There's a hidden beneficiary to encouraging clinical-trial enrolment," he says. "We also have to think about who's suffering when we delay innovation." ■



SOURCE: REF. 1

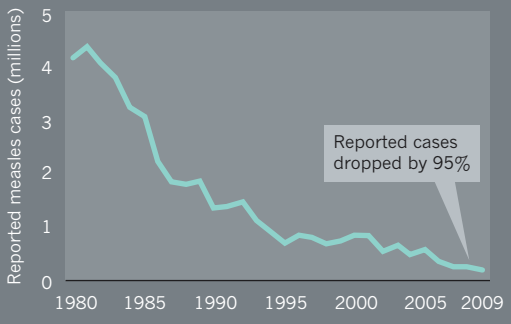
1. Malani, A. & Philipson, T. *Can Medical Progress be Sustained? Implications of the Link Between Development and Output Markets* (NBER, 2011); available at http://www.nber.org/papers/w17011.
2. Ripley, E., Macrina, F., Markowitz, M. & Gennings C. *J. Empir. Res. Hum. Res. Ethics* **5,** 57–65 (2010).

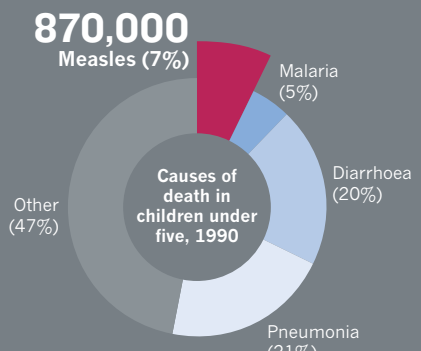
THE CASE OF MEASLES

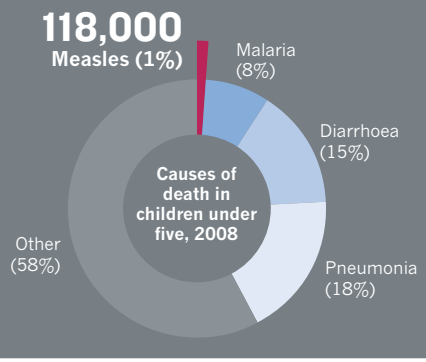
Great advances in the development and distribution of vaccines mean that some diseases can be eradicated. Measles is an important case study: efforts to stem the disease have been successful, but uneven political commitment, lack of funds and public fear threaten to undermine the progress.

PAST: A killer crushed

In 1980, before vaccination was widespread, there were around 4 million cases of measles and an estimated 2.6 million deaths from the disease worldwide[1]. Childhood mortality targets set by the United Nations, along with accelerated control programmes, have cut the proportion of childhood deaths caused by measles from 7% in 1990 to 1% in 2008 (ref. 2).

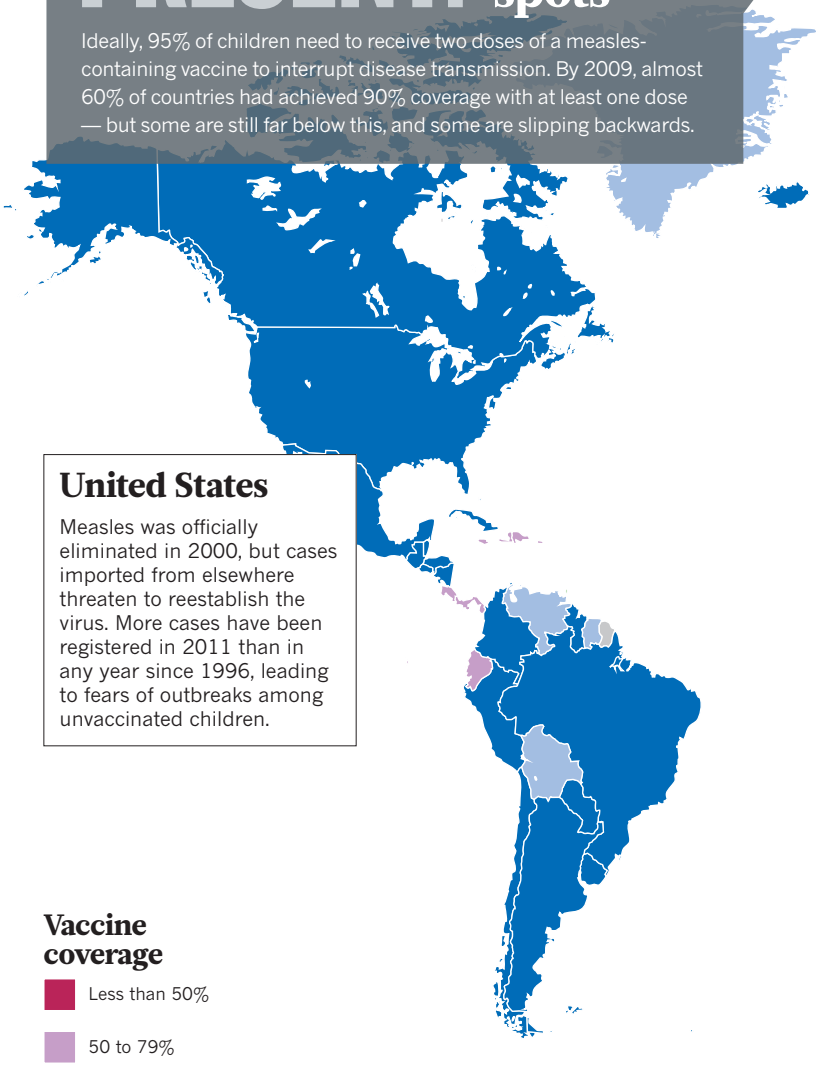






PRESENT: Trouble spots

Ideally, 95% of children need to receive two doses of a measles-containing vaccine to interrupt disease transmission. By 2009, almost 60% of countries had achieved 90% coverage with at least one dose — but some are still far below this, and some are slipping backwards.



United States

Measles was officially eliminated in 2000, but cases imported from elsewhere threaten to reestablish the virus. More cases have been registered in 2011 than in any year since 1996, leading to fears of outbreaks among unvaccinated children.

Vaccine coverage

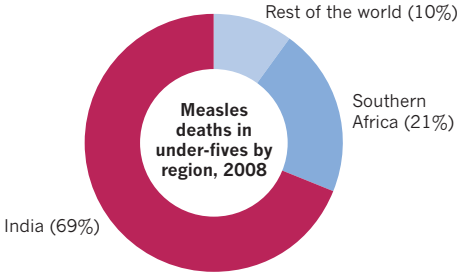
- Less than 50%
- 50 to 79%
- 80 to 89%
- 90% and over
- Unknown

Estimated coverage for the first dose of a measles-containg vaccine is provided by the World Health Organization and the United Nations' Children's Fund.

VACCINES
New promise, old doubts
nature.com/vaccines

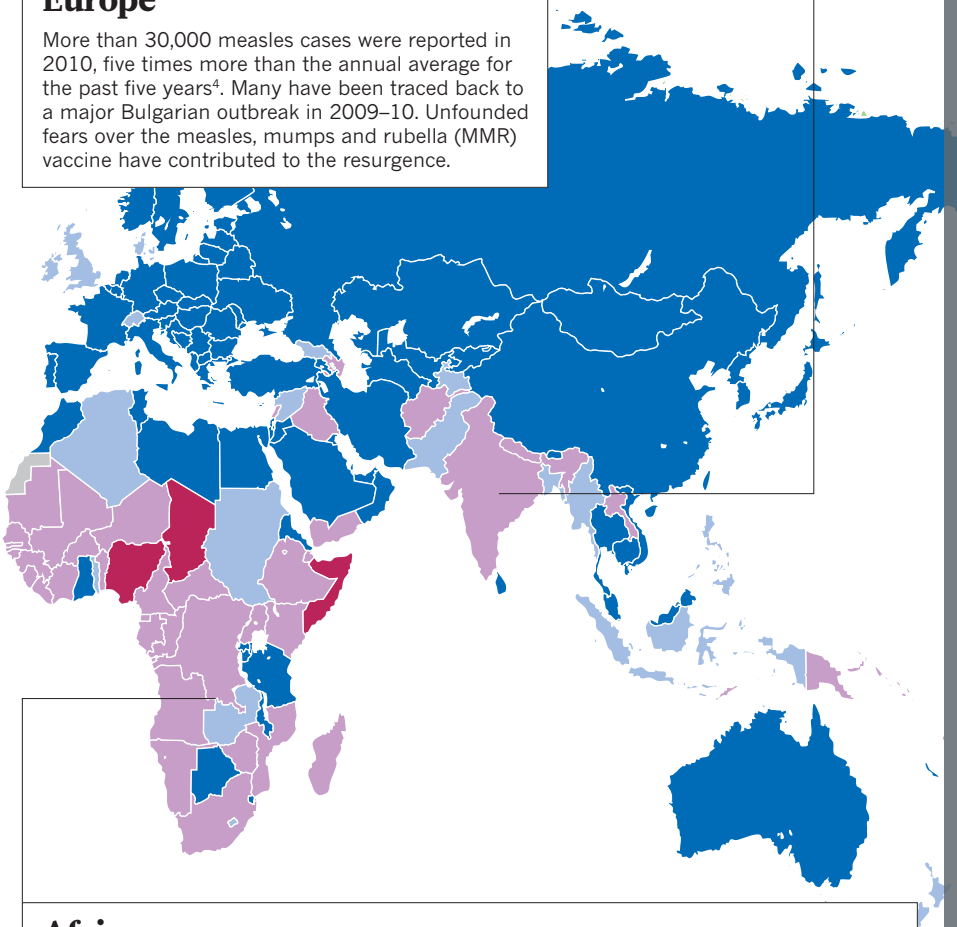
India

India is struggling to reduce deaths from measles[3], mainly because of a lack of money and political will to provide two doses of vaccine to all children. There are some indications that this is changing.



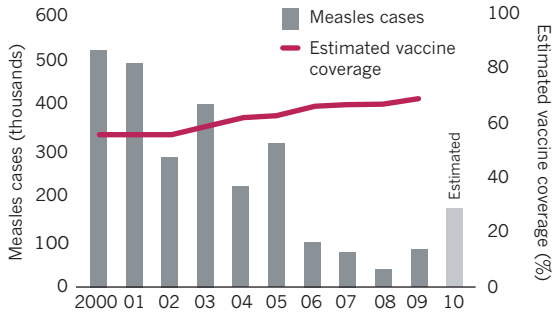
Europe

More than 30,000 measles cases were reported in 2010, five times more than the annual average for the past five years[4]. Many have been traced back to a major Bulgarian outbreak in 2009–10. Unfounded fears over the measles, mumps and rubella (MMR) vaccine have contributed to the resurgence.



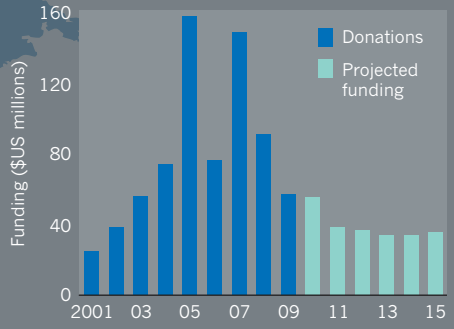
Africa

Outbreaks have been seen in 28 countries in the past two years[5], mainly because of a lack of funding and political commitment to follow-up vaccination campaigns, and problems with vaccine delivery. There has also been resistance among some religious groups in Zimbabwe, Botswana, Malawi and South Africa.

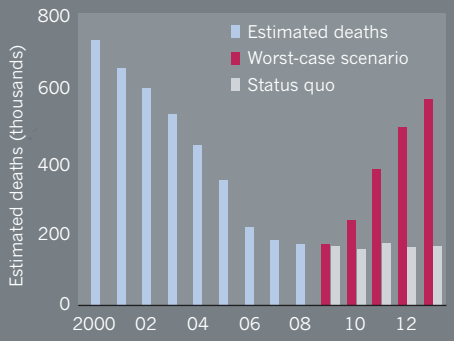


FUTURE:
Funding, fears and uncertainty

Because measles deaths have fallen, vaccination efforts now compete for funding with other diseases, so investment has dropped. Some countries are struggling to introduce the recommended second dose of measles-containing vaccine, let alone new vaccines — for invasive pneumococcal disease and rotavirus, for example — that could save many more lives.

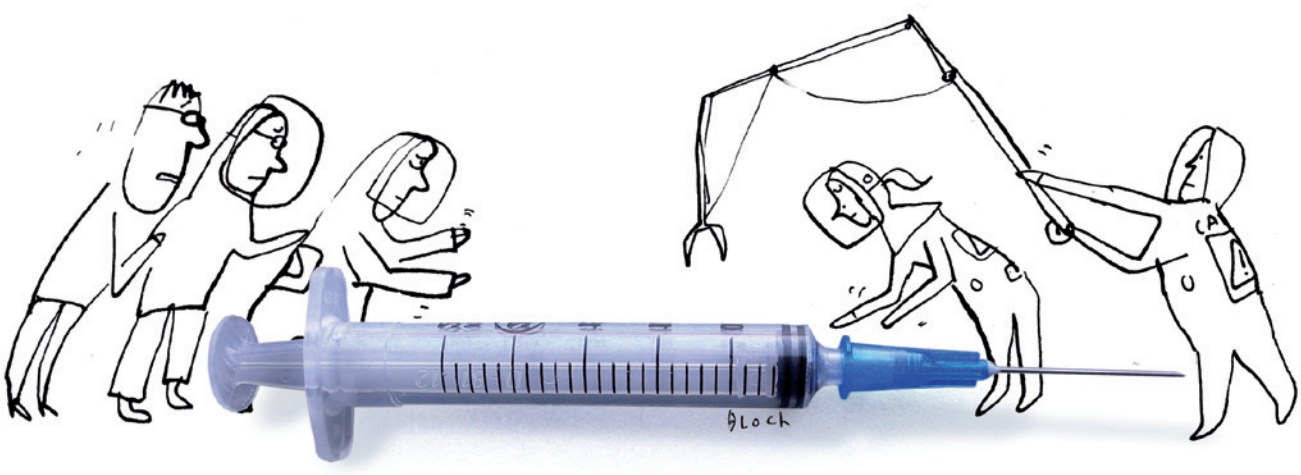


Assuming no catch-up immunizations in troubled countries, public-health officials predict a worst-case scenario in which the death toll could exceed 500,000 by 2013 (ref. 1).



In a 'status-quo' scenario, modest increases in first-dose vaccine coverage are complemented by catch-up immunizations at about 2008 levels — but this still falls short of global eradication.

1. Strebel, P. M. *et al. J. Infect. Dis.* doi:10.1093/infds/jir111 (2011).
2. van den Ent, M. V. X., Brown, D. W., Hoekstra, E. J., Christie, A. & Cochi, S. L. *J. Infect. Dis.* doi:10.1093/infds.jir081 (2011).
3. Black, R. E. *et al. Lancet* **375,** 1969–1987 (2010).
4. European Centre for Disease Prevention and Control. *Epidemiological Update on Measles in EU/EEA* (2011) available at: go.nature.com/yndhco
5. Centers for Disease Control and Prevention. *MMWR Morb. Mortal. Wkly Rep.* **60,** 374–378 (2011).



THE REAL ISSUES IN VACCINE SAFETY

Hysteria about false vaccine risks often overshadows the challenges of detecting the real ones.

BY ROBERTA KWOK

SERGE BLOCH

John Salamone is not a vaccine sceptic. He has never been persuaded by spurious claims that vaccines are toxic to children and responsible for autism or a host of other ailments. But tragically, Salamone found out first-hand that vaccines do have real, rare side effects when he saw his infant son, David, become weak and unable to crawl shortly after receiving the oral polio vaccine in 1990. After about two years of physical therapy and doctors' visits, Salamone learned that owing to a weakened immune system, David had contracted polio from the vaccine. "We basically gave him polio that day," says Salamone, who has retired from a position as a non-profit executive, and lives in Mount Holly, Virginia.

That was a known risk of the vaccination, which causes roughly one case of the disease per 2.4 million doses, often in people with an immune deficiency. A safer, inactivated, polio vaccine was available at the time, but the oral vaccine was cheaper, easier to administer and thought to be more effective at controlling outbreaks. But by the 1980s, polio had been all but eliminated in the United States; all cases originating in the country came from the vaccine. Salamone and other parents successfully campaigned for the United States to shift to the safer version in the late 1990s.

Vaccines face a tougher safety standard than most pharmaceutical products because they are given to healthy people, often children. What they stave off is unseen, and many of the diseases are now rare, with their effects forgotten. So only the risks of vaccines, low as they may be, loom in the public imagination. A backlash against vaccination, spurred by the likes of Andrew Wakefield — a UK surgeon who was struck off the medical register after making unfounded claims about the safety of the measles, mumps and rubella (MMR) vaccine — and a litany of celebrities and activists, has sometimes overshadowed scientific work to uncover real vaccine side effects. Many false links have been dispelled, including theories that the MMR vaccine and the vaccine preservative thimerosal cause autism[1]. But vaccines do carry risks, ranging from rashes or tenderness at the site of injection

VACCINES
New promise, old doubts
nature.com/vaccines

to fever-associated seizures called febrile convulsions and dangerous infections in those with compromised immune systems.

Serious problems are rare, so it is hard to prove that a vaccine causes them. Studies to confirm or debunk vaccine-associated risks can take a long time and, in the meantime, public-health officials must make difficult decisions on what to do and how to communicate with the public. Still, such work is necessary to maintain public trust, says Neal Halsey, a paediatrician at the Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland. "If we don't do the research, there will be more people who don't believe in vaccines," he says.

VICTIMS OF THEIR OWN SUCCESS

Technological advances have made modern vaccines purer and safer than their historical counterparts. Most developed countries have switched to the inactivated polio vaccine and stopped using whole-cell pertussis (whooping cough) vaccines, which are made from killed bacteria and cause relatively high rates of arm swelling, febrile convulsions and periods of limpness or unresponsiveness.

Improved safety means that researchers are sometimes searching for vanishingly small risks. Although vaccines must undergo stringent safety tests before distribution, the trials typically don't enrol enough people to catch risks on the order of one case per 10,000–100,000 people (see 'Calculating risks'). The only way to find such side effects is to deploy the vaccine in the population and watch.

Officials have become increasingly vigilant. As worries about pandemic H1N1 influenza spread in 2009–10, several companies worked to prepare as many vaccine doses as possible. Meanwhile, health officials launched an unprecedented surveillance effort to monitor the vaccines' safety. US scientists and officials studied data from voluntary adverse-event reports, managed-care organizations, health-insurance companies, immunization registries, a network of neurologists and various health-care systems. European scientists linked data from 15 countries. And Chinese officials instructed health-care workers to report potential side effects within 24 hours; for the most serious events, they had two hours.

Scientists were specifically looking for Guillain-Barré syndrome, a paralytic disorder that is often treatable but can cause long-term disability or death. A 1976 swine-flu vaccine distributed in the United States was associated with between five and nine cases per one million vaccine recipients. Studies of subsequent flu vaccines have not shown a consistent link, but officials have been on the lookout for it. During the 2009–10 pandemic, something stranger turned up: some 60 cases of narcolepsy emerged among 4- to 19-year-olds in Finland. Most had received the H1N1 vaccine Pandemrix, made by GlaxoSmithKline in Brentford, UK. Another narcolepsy cluster showed up in Sweden. Scientists have yet to confirm whether the vaccine caused the rise in incidence.

Surveillance efforts have paid off for a variety of vaccines. A rotavirus vaccine was suspended in the United States in 1999 after public-health officials received 15 reports of intussusception, an infolding of the bowel, in vaccinated infants. The mechanism is uncertain, but the live-virus vaccine might cause swelling of bowel lymph nodes and increase contraction, leading to infolding. The vaccine is estimated to have caused about one case of intussusception per 10,000 recipients.

In 2007, Nicola Klein, co-director of the Kaiser Permanente Vaccine Study Center in Oakland, California, and her colleagues found that children aged between 12 and 23 months who had been immunized with a combination vaccine for measles, mumps, rubella and varicella (MMRV) had more febrile convulsions 7–10 days after vaccination than those receiving separate MMR and varicella vaccines. The finding prompted a US immunization advisory committee to withdraw its preference for the MMRV vaccine. A subsequent study[2] suggested that the combined vaccine resulted in one more febrile convulsion per 2,300 doses than the MMR and varicella vaccines given separately.

Efforts are under way to improve surveillance in low- and middle-income countries, some of which are gaining increased access to vaccines through an international programme called the GAVI Alliance (formerly the Global Alliance for Vaccines and Immunisation), based in Geneva, Switzerland. These areas could soon see new vaccines for diseases such as dengue and cholera. In 2006, the Pan American Health Organization, based in Washington DC, started a surveillance network among five Latin American countries. The World Health Organization (WHO) in Geneva is working with 12 countries, including Iran, Tunisia, Vietnam and India, to develop methods and tools for vaccine-safety monitoring, and half are already reporting to a global database, says Patrick Zuber, the WHO's group leader of global vaccine safety.

"If we don't do the research, there will be more people who don't believe in vaccines."

Researchers have also started conducting larger clinical trials. Pre-licensure trials for two new rotavirus vaccines, RotaTeq by Merck, based in Whitehouse Station, New Jersey, and Rotarix by GlaxoSmithKline, each enrolled more than 60,000 infants to evaluate safety[3,4]. But even these large trials cannot rule out rare events, so efforts would be better spent on well planned surveillance after licensing, argues Rino Rappuoli, global head of vaccines research at Novartis Vaccines and Diagnostics in Siena, Italy. With big pre-licensure trials, "you may feel better as a regulator, but you're not answering the scientific question", he says. Preliminary post-licensure studies in Mexico have detected a possible slight increase in intussusception risk after the first dose of Rotarix, and a similar pattern has emerged in Australia for both vaccines[5]. However, some researchers speculate that rotavirus vaccination may also protect against intussusception later.

DELAYED RESULTS, LOST TRUST

Even if a possible side effect is found, long periods of uncertainty can follow. To amass convincing evidence, scientists sometimes need to do controlled studies in multiple countries, covering hundreds of thousands or even millions of people. Scientists have not yet conclusively determined whether Pandemrix contributed to the European cluster of narcolepsy cases.

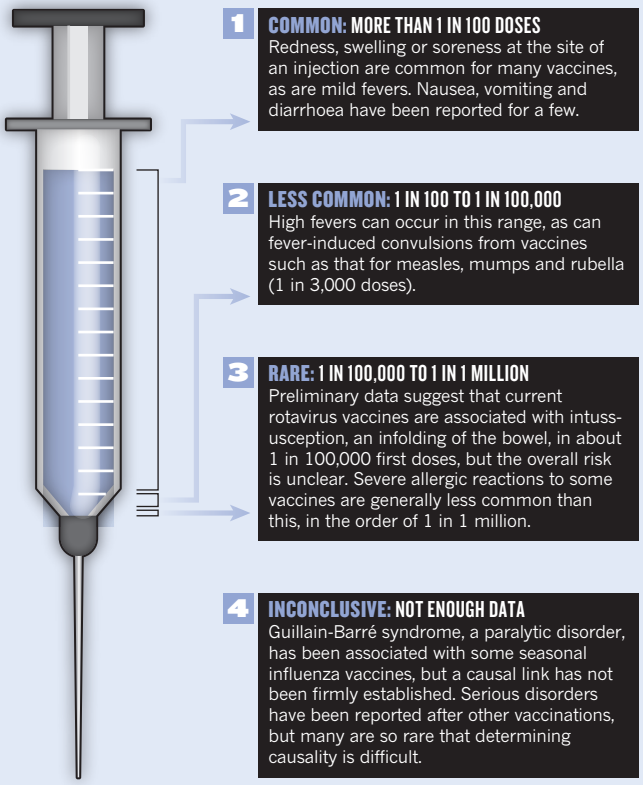
Scientists in the Vaccine Adverse Event Surveillance & Communication Consortium, a European research network, are examining narcolepsy diagnosis rates and comparing cases with matched controls across several European Union countries, some of which used different H1N1 vaccines. Data suggest that diagnosis rates rose slightly in several countries starting in 2008, before H1N1 vaccines were being distributed, but not enough to explain the episode in Finland, says principal investigator Miriam Sturkenboom, a pharmacoepidemiologist at Erasmus University Medical Center Rotterdam in the Netherlands. GlaxoSmithKline is also funding a study in Canada, where an H1N1 vaccine nearly identical to Pandemrix was used, but no rise in narcolepsy has been reported.

The increase in narcolepsy diagnoses might be explained by heightened disease awareness or infections with the H1N1 virus itself, says Jan Bonhoeffer, a paediatric-infectious-disease specialist at the University Children's Hospital Basel in Switzerland, and chief executive of the Brighton Collaboration, an international vaccine-safety research network. He says that the narcolepsy story fits a familiar pattern, similar to that seen with MMR and autism: people are eager to find an underlying cause for a serious, chronic, poorly understood disease.

Researchers need to investigate possible safety issues quickly, Bonhoeffer adds. Otherwise, by the time scientists conclude that a concern is unfounded, "no one cares, and it takes years to build up the trust again", he says. "So often, the widely communicated concern has caused more harm than it intended to prevent." A global vaccine-safety network would give scientists a faster way to test hypotheses with sufficient sample sizes, he says. In that spirit, the WHO is coordinating a global study on pandemic H1N1 flu vaccines and Guillain-Barré syndrome.

CALCULATING RISKS

Some vaccines have risks that are common but mild. A few have more serious risks, but these are very rare.



1 COMMON: MORE THAN 1 IN 100 DOSES
Redness, swelling or soreness at the site of an injection are common for many vaccines, as are mild fevers. Nausea, vomiting and diarrhoea have been reported for a few.

2 LESS COMMON: 1 IN 100 TO 1 IN 100,000
High fevers can occur in this range, as can fever-induced convulsions from vaccines such as that for measles, mumps and rubella (1 in 3,000 doses).

3 RARE: 1 IN 100,000 TO 1 IN 1 MILLION
Preliminary data suggest that current rotavirus vaccines are associated with intussusception, an infolding of the bowel, in about 1 in 100,000 first doses, but the overall risk is unclear. Severe allergic reactions to some vaccines are generally less common than this, in the order of 1 in 1 million.

4 INCONCLUSIVE: NOT ENOUGH DATA
Guillain-Barré syndrome, a paralytic disorder, has been associated with some seasonal influenza vaccines, but a causal link has not been firmly established. Serious disorders have been reported after other vaccinations, but many are so rare that determining causality is difficult.

Source: US Centers for Disease Control and Prevention. For more information, see go.nature.com/s7rfio

But strictly controlled randomized trials — the highest standard of evidence for determining causality — are often not possible because of the large number of participants needed. And randomized trials in one location will not prevent some researchers questioning whether the results apply in others, says Alfred Berg, a clinical epidemiologist at the University of Washington in Seattle.

Even if surveillance efforts became faster and more thorough, public-health officials still need to make quick decisions with incomplete data. Authorities often err on the side of caution, but warnings can make the public wary. In March, for example, Japanese officials suspended a vaccine for pneumococcal illnesses and one for *Haemophilus influenzae* type b when four children died shortly after immunization. Officials later concluded that there was no direct evidence of a link, but the episode still caused a scare, says Pier Luigi Lopalco, head of the vaccine-preventable-diseases programme at the European Centre for Disease Prevention and Control in Solna, Sweden. Suspending a vaccine tends to get more media attention than resuming one, he says, so people remember only the threat.

US government officials have drawn criticism for pushing for removal of thimerosal from vaccines, despite a lack of evidence that it poses a risk. "People said, why are you removing this if it's not a problem?" says Ken Bromberg, a paediatrician at the Brooklyn Hospital Center in New York. "It must really be a problem even though you say it's not." But inaction would have caused a loss of credibility, says Halsey. "That is not something I think the public would have accepted."

FINDING THOSE IN DANGER

Researchers have long known that some individuals are more susceptible to vaccine risks than others. Immunocompromised individuals have generally been discouraged from receiving live-virus vaccines. But other possible vulnerabilities are less clear. Some speculate that children with metabolic disorders might be prone to vaccine side effects, but two studies published in April suggest otherwise. Klein and her colleagues reported[6] that children with inherited metabolic disorders do not show an increase in emergency-department visits or hospitalizations in the 30 days after being immunized. The other study found that children with one type of metabolic disorder — urea cycle disorders — did not have more serious metabolic problems than usual within 21 days of vaccination[7].

Some researchers hope that doctors will eventually be able to screen people for genetic predispositions to vaccine side effects. Gregory Poland, a vaccinologist at the Mayo Clinic in Rochester, Minnesota, says that once predispositions have been identified, genetic screening would at least make the risks and benefits explicit. Scientists have begun studying predispositions to side effects from smallpox vaccination: Kathryn Edwards, a vaccinologist at Vanderbilt University in Nashville, Tennessee, and her colleagues have reported[8] two genes that might be associated with reactions such as rashes, and Poland's team is searching for genetic risk factors for myopericarditis — inflammation of the heart muscle and surrounding tissue.

Even if immunization does prove risky for certain children, withholding the vaccine could pose a greater threat. Vaccine-preventable diseases can be particularly severe or even fatal for patients with metabolic disorders, says Marshall Summar, chief of the division of genetics and metabolism at the Children's National Medical Center in Washington DC.

Edwards and her colleagues have been studying how children with mitochondrial disorders, a group of metabolic disorders, respond to vaccines and natural infections. If vaccines present a risk, doctors could take steps to counteract possible effects, for example by ensuring that the child is well nourished after immunization, says Edwards.

Safer vaccines and manufacturing processes are also in the works. A Novartis plant in Holly Springs, North Carolina, will produce influenza vaccine doses in cell culture, rather than the industry-standard chicken eggs. This process will improve reliability and reduce allergic reactions to egg proteins, says Rappuoli. The plant will be ready to make pandemic-flu vaccine this year if needed, he says.

Researchers are also developing replacements for vaccines that can be risky for vulnerable groups. These include current smallpox vaccines that cannot safely be given to immunocompromised people; the tuberculosis vaccine, which is not recommended for HIV-positive infants; and the yellow-fever vaccine, which puts elderly people at particular risk of a yellow-fever-like illness. The challenge will be to make safer vaccines just as effective: James Cherry, a paediatric-infectious-disease specialist at the University of California, Los Angeles, speculates that an outbreak of whooping cough in California in 2010 might have occurred partly because the safer acellular pertussis vaccines now in common use in developed countries tend to be less effective than the best whole-cell vaccines.

Researchers are quick to emphasize that the benefits of vaccines still greatly outweigh the risks. But as diseases recede from the public's memory, the population's tolerance for side effects will drop even further. "If you don't know the diseases and you haven't seen them, then you really aren't willing to accept any risk," says Edwards. Despite scientists' best efforts, eliminating risk is impossible. Vaccines are biological products with biological effects, says Juhani Eskola, deputy director general of Finland's National Institute for Health and Welfare in Helsinki. "We can never make them 100% safe." ■ SEE EDITORIAL P.420

Roberta Kwok *is a freelance writer in Burlingame, California.*

1. Immunization Safety Review Committee *Immunization Safety Review: Vaccines and Autism* (National Academies Press, 2004).
2. Klein, N. P. *et al. Pediatrics* **126,** e1–e8 (2010).
3. Ruiz-Palacios, G. M. *et al. N. Engl. J. Med.* **354,** 11–22 (2006).
4. Vesikari, T. *et al. N. Engl. J. Med.* **354,** 23–33 (2006).
5. Buttery, J. P. *et al. Vaccine* **29,** 3061–3066 (2011).
6. Klein, N. P. *et al. Pediatrics* **127,** e1139–e1146 (2011).
7. Morgan, T. M. *et al. Pediatrics* **127,** e1147–e1153 (2011).
8. Reif, D. M. *et al. J. Infect. Dis.* **198,** 16–22 (2008).

HIS BEST SHOT

BY CORIE LOK

Can Bruce Walker transform HIV vaccine research?

Bruce Walker didn't want to sit next to Terry Ragon on the 24-hour plane ride from Boston to South Africa. He had only recently met the wealthy, Cambridge, Massachusetts-based software executive and was about to spend two full days touring AIDS-ravaged Durban with him in hope of obtaining a donation. Walker, an immunologist and physician at Massachusetts General Hospital (MGH), wanted to give Ragon some space and get some work done, but Ragon insisted they sit together. During the flight, he peppered Walker with questions about his research in South Africa. He also warned him not to get his hopes up. "I go on a lot of these kinds of trips, and I don't give people very much money," Ragon said.

Walker was disappointed, but he stuck to the plan. He took Ragon to the crumbling, 100-year-old McCord Hospital, where he followed doctors and visited impoverished, young people with HIV. "All three of the patients I sat in with were going to die, and one of them was dying right there in front of me," says Ragon. He had been to Africa before but never had he so intimately seen the pain and suffering caused by AIDS.

As the trip neared its end, Walker knew that it was time to broach the subject of money ▶

VACCINES
New promise, old doubts
nature.com/vaccines

T. GRAY/GETTY

again. He had been trained by MGH fundraisers to give potential donors a range of options. For a modest sum, US$5,000–20,000, Ragon could fund lab equipment or nurses — $1 million might fund a small clinical trial. But on a whim, Walker decided to float a more ambitious idea: creating an institute in which researchers from different fields could focus solely on HIV vaccines under one roof, with the kind of funding that would enable high-risk projects. "I thought the idea was half-baked," says Ragon, "but it intrigued me."

That was in March 2007. Talks continued, and about a year later Ragon and his wife, Susan, agreed to give $100 million over 10 years to create the Ragon Institute of MGH, MIT and Harvard. Established in early 2009, with Walker as its director, the institute was a positive note at a challenging time for HIV vaccine development. In late 2007, the pharmaceutical company Merck announced that a high-profile vaccine candidate in a large phase II clinical trial failed to protect people from infection with HIV, and even increased the risk of infection for some[1]. Then, later in 2009, another surprise came with the results from a huge phase III trial in Thailand, dubbed the RV144 trial, showing that a combination of two previously unsuccessful vaccines had defied expectations and provided modest protection from infection[2]. Although researchers were cautious about the work, the results offered a glimmer of hope that protection was possible.

"Walker provides an environment that inspires innovation."

Researchers say that the trials — two of only three major vaccine efficacy studies undertaken in 25 years of research — have reminded them of just how little they know about harnessing the immune system to block this deadly disease. Researchers have called for a return to basic HIV research (see 'Where the money goes'). But they need stronger interdisciplinary collaboration, innovative trial designs and, perhaps most of all, freedom — through funding — to take chances. Now, many say that the Ragon Institute has the opportunity to put these factors to work. Walker can clearly draw the money, and Anthony Fauci, director of the National Institute of Allergy and Infectious Diseases in Bethesda, Maryland, says that Walker can also gather the right team. "He has a special talent for getting groups of people together from diverse backgrounds in a collaborative, synergistic way," he says. "The challenge," says Herbert Virgin, an immunologist at Washington University in St Louis, Missouri, and head of the institute's external scientific advisory board, "is now putting it all together to generate a vaccine."

HOOKED ON BASICS

Two months ago in his MGH office, Walker was hanging on the words of Krista Dong, a Ragon Institute physician who lives in South Africa. Dong was rifling through detailed, handwritten notes about a clinical study she has been helping to design. The project aims to look at the immunological events during the first few days of HIV infection — most studies haven't been designed to capture this information. The researchers planned to collect blood samples from 200 young, uninfected women twice a week for a year.

The research could provide crucial information about how the virus takes hold in the body — information that cannot be gleaned easily from animal studies — but it requires extraordinary cooperation and trust from the participants. As Dong explained how she and her team would do this through HIV-prevention and job-training programmes, the smile on Walker's face grew. Walker is a listener and, unlike many of his Boston peers, he speaks softly and slowly. Walker asked a few well chosen questions on logistics but was clearly sold. "Let's start!" he said, then began to list possible sources of funding, mostly from foundations and philanthropists. Although he hadn't actually secured the money, he urged Dong and her team to plough ahead. "Bruce is endlessly optimistic in an infectious way," says Dong. "He provides an environment that inspires innovation and personal drive."

Walker's education as a fundraiser began in the mid 1990s, when he was head of the Partners AIDS Research Center at the MGH. He learned from a development officer there how to ask for a million dollars, something Walker found difficult at first. "How could I ask that?", he recalls thinking. "But then I realized — how could I, in good conscience, not ask for help?"

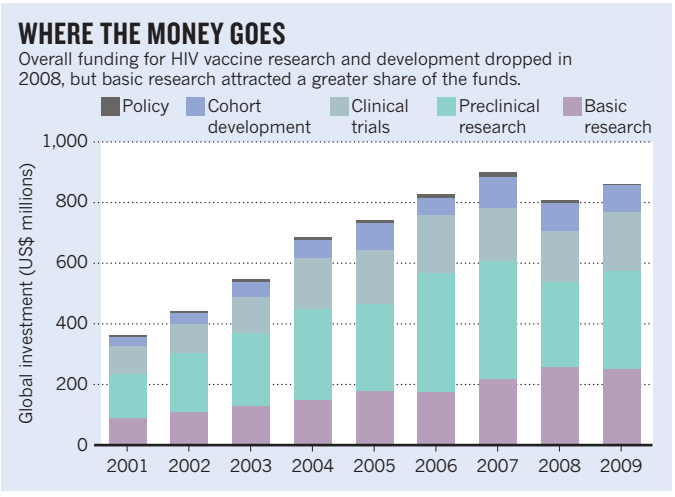
He soon learned the power of showing, rather than telling, people what their money can do. In 2000, a postdoctoral fellow working out of a closet-sized lab in Durban encouraged Walker to visit what was becoming the epicentre of the African AIDS epidemic. South Africa had come to lead sub-Saharan Africa in both the number of people living with HIV (5.6 million in 2009) and the number dying from AIDS (310,000 in 2009). Durban is the largest city in the most highly affected province, KwaZulu-Natal. Here, six in ten women are HIV-positive by the age of 23.

Walker was moved by what he saw. And when he met with other local HIV researchers looking for a bigger, better lab space in which they could work together, they hatched an idea to build a new research institute. Walker sent a proposal to the Doris Duke Charitable Foundation in New York, which was already funding some of his work. As he would do for Ragon several years later, he invited the head of the science programme to Durban. Shortly afterwards, the foundation committed $1.8 million to the construction of a new building, and another $2.25 million to support research and training for four years. The Doris Duke Medical Research Institute opened in 2003 on the campus of the Nelson R. Mandela School of Medicine at the University of KwaZulu-Natal. It was the foundation's first major international grant for HIV. Walker maintains close ties with the researchers at the institute through frequent phone calls and videoconferences. He also makes the long flight over there every other month, and co-supervises two PhD students and a postdoc.

Walker's laid-back, affable style helps him connect with people, and his connections have greatly helped his work. With his collaborators, he developed a cohort of around 1,200 people with HIV in South Africa. Studies on these individuals have provided many insights over the past decade, demonstrating, for example, how the virus evolves as the disease progresses. "It's not easy establishing all those links and collaborations and trust," says Andrew McMichael, an HIV immunologist at the University of Oxford, UK.

Walker has also taken an interest in the roughly 1 in 300 people with HIV who are able to keep the virus in check without any drugs and don't progress to AIDS. Walker and his colleagues painstakingly tracked down some of these rare patients, known in some circles as 'elite controllers', by connecting with HIV patient groups and physicians. They built up a cohort of about 1,500 controller patients, along with a bank of their blood samples, which they have shared with other research groups. They want to discover how these people suppress the virus, in the hope that the mechanism can be mimicked using a therapeutic vaccine.

The clues are mounting. A type of immune-system cell called a $CD8^+$



P. WEINBERG / HHMI

Bruce Walker takes potential funders to meet patients and staff such as Nurse Kesia Ngwenya at McCord Hospital in Durban, South Africa.

or 'killer' T cell may exert selective pressure that allows only weaker versions of the virus to survive in controllers[3]. "Our direction right now is to try to understand how exactly these cells are driving these viruses to be less fit," says Walker. But some researchers think that the controllers' killer T cells are able to reproduce more and produce larger amounts of perforin — a protein that can poke holes in infected cells to help kill them[4]. Whatever mechanisms are at work, exploiting them will be a challenge, says Larry Corey, the principal investigator for the HIV Vaccine Trials Network (HVTN) and the head of the Fred Hutchinson Cancer Research Centre in Seattle, Washington. Studying the controllers "is important conceptual work", he says. "How to translate that into making an effective vaccine is easier said than done."

BRANCHING OUT

Meanwhile, Walker has been trying to pull researchers from other fields into the fold. From the start, he wanted to build an interdisciplinary team to oversee the Ragon Institute, one that could bring fresh perspectives to bear on issues that have dogged HIV researchers for 25 years. Walker recruited a steering committee to help him oversee the institute, including a materials scientist and a computational biologist, both from the Massachusetts Institute of Technology (MIT) in Cambridge. A group of 14 labs at the MGH form the core of institute, which funds collaborative research projects headed up by at least two principal investigators. One key area Walker has built up in his institute is basic HIV research, which has been held back by inadequate animal models. The HIV field, he says, has grown insular, with little interaction with immunologists doing basic research. He brought in a leading immunologist, Laurie Glimcher at the Harvard School of Public Health in Boston, Massachusetts. Glimcher had a history of branching out into different disciplines but had never worked on HIV. She now heads the institute's basic immunology programme and is overseeing the development of a humanized mouse core — a bank of mouse models that have key components of a human immune system. Walker hopes that these mice will allow researchers to test preliminary vaccines *in vivo* sooner than they can now.

Walker also pushed to have physical scientists join the team, something that resonated with Ragon, who has a physics degree from MIT. Three years ago, Walker approached Arup Chakraborty, a computational immunologist at MIT who had not studied HIV. Chakraborty was sceptical that he could contribute much to the field. But, as for others, an emotional trip to South Africa changed his mind. Now he heads the computational biology programme at the institute and more than one-third of his lab's work is devoted to HIV. Chakraborty's modelling expertise has helped to reveal important information in the wealth of immunological data already available. His preliminary work has shown, for example, that some variations in the HIV genome are linked. This means that if some mutations happen without others, the virus might become vulnerable to immune attack, says Walker.

Walker's overarching goal for the institute is to contribute to the development of an HIV vaccine and, while doing so, create a team culture rather than one that emphasizes individual credit. "It is going to take all of us making a contribution," says Walker.

Walker says that better modelling of data and broader collaborations could feed into a new breed of clinical trial. Typical phase II and III vaccine trials are geared towards showing efficacy. That means that they need to recruit enough patients to be able to statistically show an effect. For a single trial, that can take years and cost many millions of dollars. (The RV144 trial enrolled more than 16,000 participants at a cost of $103 million, which is typical of vaccine phase III trials.)

Walker aims to do a different type of vaccine trial: one that enrols 10–20 people, half of whom would receive a placebo. Researchers would then do a detailed analysis of the patients' immune responses, from T-cell activities to antibody generation. He hopes that the analyses, when done alongside or even before larger efficacy trials, would provide more clues about whether and how a vaccine candidate is stimulating the immune system — a crucial missing piece of the HIV vaccine puzzle. Such smaller, faster trials would allow candidates to be tested in parallel and hopefully give quicker indications of success or failure for less cost. This approach has been tried before, but until recently few HIV vaccine candidates elicited a strong enough response to study in this way.

Two vaccine candidates inherited by the institute are now in phase I trials; one of them is being tested in collaboration with HVTN and the International AIDS Vaccine Initiative (IAVI), headquartered in New York. The institute is also developing other candidates and is collaborating with the IAVI to build clinical laboratory infrastructure in Durban, where testing can be done more cheaply than in Western countries. The facility should be operating by early 2013.

Getting vaccine candidates into clinical trials, and testing them in innovative ways at a lower cost, will be key to the success of the Ragon Institute. And to encourage more interactions between the groups, Walker's lab, along with several other Ragon labs at the MGH and MIT plan to move in about a year into a new building near MIT. Ragon researchers based at other institutions will have space there as well.

But future trials will depend on whether Walker can continue to raise funds. Even $100 million, the largest donation in the MGH's history, is not enough to tackle the sheer complexity of the virus and the large number of unknowns about how to stimulate the immune system to fight off the disease, Walker says. He is still searching for funding from donors and chasing grants. He will undoubtedly be bringing more people to Durban. And he remains optimistic. "This is a solvable problem," he says. "There's no time to waste." ■

Corie Lok *is* Nature*'s Research Highlights editor in Cambridge, Massachusetts.*

1. Buchbinder, S. P. *et al. Lancet* **372,** 1881–1893 (2008).
2. Rerks-Ngarm, S. *et al. N. Engl. J. Med.* **361,** 2209–2220 (2009).
3. Miura, T. *et al. J. Virol.* **83,** 2743–2755 (2009).
4. Hersperger, A. R., Migueles, S. A., Betts, M. R. & Connors, M. *Curr. Opin. HIV AIDS* **6,** 169–173 (2011).

COMMENT

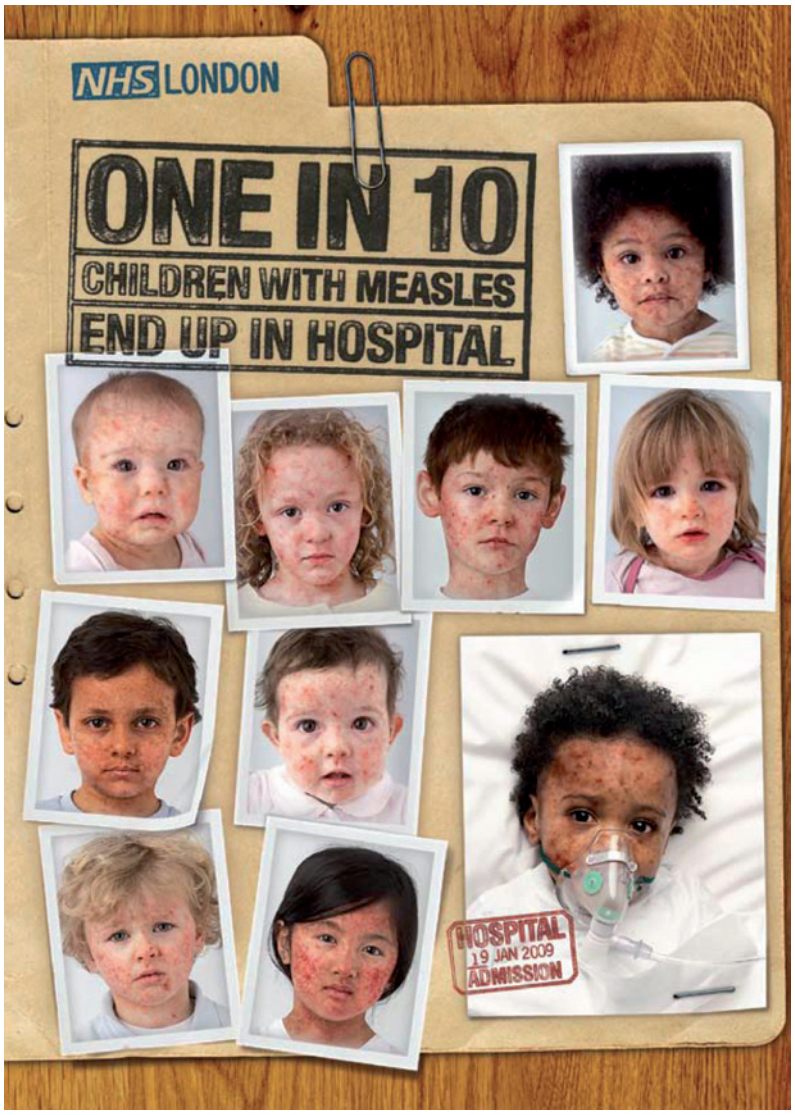
VACCINES Lessons from the long war to eradicate polio **p.446**

RENEWABLE ENERGY Lithium feeds the chase for a superbattery **p.448**

PSYCHOLOGY How jokes reveal the workings of the mind **p.450**

ART EXHIBIT Visions of the future for the human body **p.451**

NHS LONDON



Are campaigns like this the best way to reach doubting parents?

Target the fence-sitters

Past waves of vaccine rejection in industrialized nations have a lot to teach us about preventing future ones, argues **Julie Leask**.

Immunizing a child requires a leap of faith by any parent or carer. Picture Emily, a new mother, whose healthy eight-week-old baby is scheduled to receive vaccines against up to eight diseases that Emily has never seen. Emily feels wary of expert knowledge. She is concerned that the vaccines could weaken her baby's immune system and is anxious about the technologies of modern life. Prosaically, she feels daunted by the trip to a clinic full of sick people where there might not be anywhere to change or feed her baby comfortably.

Emily seeks information online. Three of the first ten search results link vaccines to problems such as allergies, autism, diabetes and cancer. One might expect Emily and many other new parents in industrialized countries to be rejecting immunization.

Surprisingly, levels of support for childhood vaccinations are generally high and stable. In countries that are members of the Organisation for Economic Co-operation and Development, 95% of children, on average, received all three primary doses of diphtheria-tetanus-pertussis (DTP) vaccine in 2009. The United Kingdom's measles-mumps-rubella (MMR) immunization rates have clawed back to 89% from a 2004 low of 80% that was caused by the now debunked claims of a link to autism[1] (see 'The cost of a scare'). Coverage for other vaccines was unaffected. The United States has recorded 95% DTP immunization rates for toddlers and rates of children receiving no vaccines remain stable at four to six per thousand[2]. Australia's immunization rates have increased steadily over the past decade to 92.5% of two-year-olds fully vaccinated in 2008[3]. Five out of the six World Health Organization regions achieved a 90% reduction in measles deaths between 2000 and 2010. Finland, Cuba, England and Wales, Brazil, Mexico, the United States, Canada, South Korea and Australia are at, or near, measles elimination.

CLIMATE OF DISTRUST

But dig a little deeper and there are grounds for concern. From 2008 to 2009, the United States recorded a 3% decline in MMR ▶

VACCINES
New promise, old doubts
nature.com/vaccines

immunization rates to 90.6%[4]. In the United States, Britain and Australia, up to one-third of parents report concerns about the number of vaccines that babies are now receiving and are more distrustful of newer vaccines, a phenomenon that may increase as more are introduced. A survey in 2008–09 found that up to about one–fifth of parents from five European countries reported doubts about having their child vaccinated[5].

Communities of parents — particularly those who espouse the alternative lifestyles, anthroposophical or religious beliefs that oppose vaccination — continue to contribute to outbreaks of diseases such as pertussis, measles and *Haemophilus influenzae* type b (Hib).

But the greatest causes for concern are unfounded scares around particular vaccines — leading to anything from small downturns in immunization rates to the cessation of entire programmes. Japan had one such scare in the 1970s, when the deaths of two children within 24 hours of receiving the DTP vaccine led to the suspension of that programme and then its resumption two months later with a primary dose beginning at two years of age rather than at three months. A pertussis epidemic followed in 1979 with more than 13,000 cases leading to 41 deaths[6]. Britain's recent MMR experience pales in comparison with its own DTP scare of the late 1970s when the vaccine was linked to encephalopathy. Immunization rates fell from 80% to 30%; there followed 70 deaths among more than 300,000 cases of pertussis[7].

As a social scientist specializing in immunization take-up, it is clear to me that we can and must work harder to head off such scares by better engaging fence-sitting parents and wavering health professionals. Just as vaccine programmes must be informed by sound research, so too must communication strategies.

SCARE STORIES

What makes a vaccine scare take hold? It is a complex interplay of factors embedded in a country's historical, social and political context.

Sporadic media reports do not immediately affect vaccine uptake. The media tends to sideline groups that criticize vaccines until a prominent figure champions a theory against a backdrop of mistrust in government. The MMR–autism and DTP–encephalopathy links were advanced by doctors committed to their hypotheses; both were charismatic individuals from respected institutions laying their theories Galileo-like at the door of the scientific church.

The British doctor Andrew Wakefield, who linked the MMR vaccine with autism, juxtaposed stereotypes of hard-pressed parents and kindly clinicians against those of unyielding health authorities. His views fed a hunger for autism's cause. A similar hunger drove the now equally discredited attempt to link the DTP vaccine with sudden infant death syndrome (SIDS), which lost its currency by the late 1990s with better understanding of the precursors of SIDS.

Such figures can give a scare enough traction or politicization to become mainstream. At this point, media editors often marginalize medical reporters who are knowledgeable enough to discern quality of evidence in favour of general newshounds. Reports may then begin to give weight or 'false equivalence' to theories with scant scientific support.

Health professionals are key in tipping a scare towards widespread vaccine rejection. They too are affected by persistent public messages. In 1976, at the height of Britain's DTP scare, up to one-third of general practitioners (GPs) were advising against pertussis immunization[8]. In 1998, just four months after the publication that triggered the MMR scare, 13% of GPs and 27% of practice nurses in north Wales thought it very likely or possible that the vaccine was associated with autism[9]. Committed, confident and knowledgeable health professionals are the cornerstone of successful immunization programmes. Parents repeatedly rate them as their most trusted source of advice. So if doctors and nurses lose confidence, it can have a profound effect.

Many commentators assume that a failure to vaccinate is caused by parents' poor understanding of immunization. Under this logic, parents who are given scientific facts will abandon their erroneous beliefs and proceed to vaccinate. However, the work of Nobel laureate Daniel Kahneman and Amos Tversky and others on heuristics and biases demolished these assumptions. Decisions about whether to immunize are not usually made rationally nor at one moment in time. And knowledge rarely predicts vaccine uptake — indeed, refusers are more likely to have university education than those who accept vaccination. Hence scientific arguments alone will not sway them, and may even increase their resolve to not immunize.

THREE STEPS

There are three ways in which governments can maintain or retain high uptake of safe, effective childhood vaccines. First, they must minimize the structural barriers. For every concerned parent or carer like Emily, there is another who finds it difficult to get her child immunized on time because of practical barriers such as a lack of transport, money or help to mind other children.

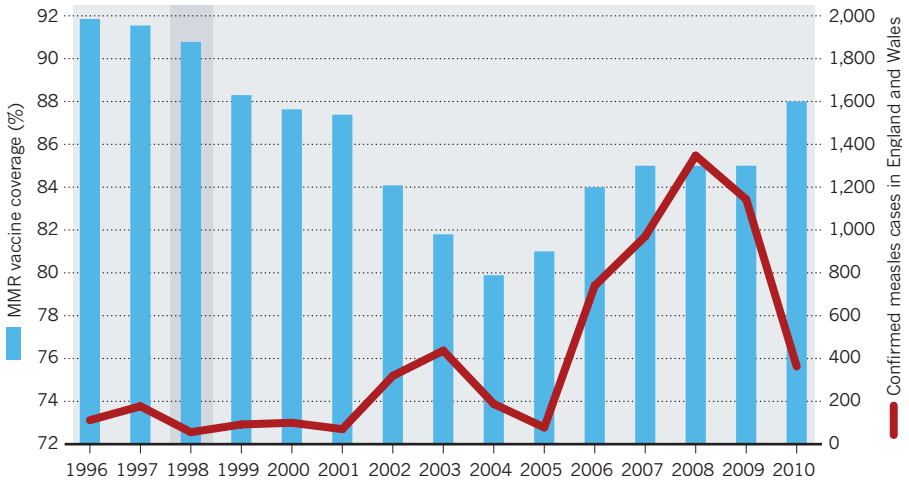
Countries with high child-immunization rates have well-oiled systems: free and accessible vaccines, national record keeping and reminders. Financial incentives for parents and providers and sanctions such as exclusion of unvaccinated children from childcare during outbreaks or compulsory

T. IRELAND/PA ARCHIVE/PRESS ASSOCIATION IMAGES

UK doctor Andrew Wakefield retained support when he was struck off the medical register in May 2010.

THE COST OF A SCARE

In the wake of the now-debunked claims in 1998 of a link between the measles, mumps and rubella vaccine and autism, vaccination dropped and measles cases rose in England and Wales.



SOURCE: HEALTH PROTECTION AGENCY

immunization also have an effect. But no intervention works in isolation and programmes must be comprehensive to succeed.

Second, communication strategies need to be tailored to groups for whom real gains can be made. Between 3% and 7% of all children are under-vaccinated because their parents refuse some or all vaccines; these parents tend to have intractable views. Hesitant parents such as Emily are a larger and more attentive group who usually vaccinate but might delay or decline a stigmatized vaccine.

Communication with this group should be the priority and needs to be informed by better evidence. Governments and health organizations must move beyond deficit models of communication that assume the public to be passively awaiting their information fill. Rather, they must recognize that people interact with information according to their experiences and social settings[10].

Tools can include: motivational interviewing — where health professionals guide hesitant parents to engage with the issue and elicit motivation for change while respecting their autonomy; decision aids (such as that of Australia's National Centre for Immunisation Research and Surveillance; go.nature.com/hp6l56) that help parents to consider the pros and cons of their options; peer-led and expert-resourced parent discussion groups; and social-media strategies that address rumours and promote vaccination.

Third, health professionals must be kept on board. This involves initiatives to sustain their confidence in safe vaccines and raise their competence to address parental concerns. More time should be spent on immunization in medical and nursing curricula; continuing education should be provided; and timely updates issued when scares arise. More pragmatically, systems should be put in place to prompt doctors or nurses when a vaccine is due or overdue, to evaluate their performance as vaccination providers, and to enable suitably qualified health professionals to give a vaccine without a doctor's involvement each time.

Better government engagement of health professionals and the public will also enhance systems for reporting and acting on adverse events following immunization. An atmosphere that censors any public concerns can unwittingly hinder efforts to hear and respond to real problems and can alienate hesitant parents, the most important audience to keep on side.

> *"An atmosphere that censors any public concerns can unwittingly alienate hesitant parents."*

In sum, anti-vaccine sentiment is inevitable, so the professionals involved should be prepared. It is too late once a scare arrives. Countries need to monitor and engage with their public and professionals and develop communication plans pro-actively. In that communication they also need to assure the public that a truly robust programme of proactive research continues to explore the safety of existing and emerging vaccines. The United States has led the way, for example, in holding workshops with the public that informed the government's vaccine safety research agenda.

THE FUTURE

Many questions remain about the precursors to large declines in vaccine acceptance. The UK and US governments have ongoing surveys to measure attitudinal trends. Other governments should commit to similar evaluations both of coverage and of public attitudes, and surveys should be harmonized for comparison across countries and over time[5]. Furthermore, researchers should ground their studies in theories of health behaviour and use validated measures. Such measurement needs to be complemented by qualitative enquiry, asking the 'why' and 'how' questions. For example, interviews with new parents could explore how they negotiate anti-vaccination information from their social-media networks.

The MOTIV (Motors of Trust in Vaccination) Think Tank initiated by Sanofi Pasteur and the London School of Hygiene and Tropical Medicine was established in December last year to better understand the diverse factors that drive immunization rates. This multidisciplinary group has proposed a research agenda centred on three broad areas: decision-making, social norms and communication. Questions include: what cognitive processes underpin vaccine decision-making and what are their relative weights in different contexts? How do social networks shape disease and vaccine perceptions? How does public engagement influence levels of trust in vaccines and vaccination-promoting groups or organizations? The group is launching an international Centre for Decision-Making on Immunisation to take forward multidisciplinary research to address these questions.

The safest and most effective vaccines are of little use if too few people take them. Public support for immunization remains high in most industrialized countries, but vaccine scares will continue. Our strategies must be tailored to our times — they must be consultative and grounded in sociology, psychology and communication science. ■

Julie Leask *is at the National Centre for Immunisation Research and Surveillance, Discipline of Paediatrics and Child Health, School of Public Health, University of Sydney, New South Wales 2006, Australia.*
e-mail: JulieL3@chw.edu.au

1. Quarterly Vaccination Coverage Statistics for Children Aged up to Five Years in the UK (COVER programme): July to September 2010 in *Health Protection Report* **4**(38) (Health Protection Agency, 2010).
2. Centers for Disease Control and Prevention. *Morb. Mortal. Wkly Rep.* **59,** 1171–1177 (2010).
3. Hull, B. P., Mahajan, D., Dey, A., Menzies, R. I. & McIntyre, P. B. *Commun. Dis. Intell.* **34,** 241–258 (2010).
4. National Committee for Quality Assurance. *The State of Health Care Quality* (NCQA, 2010).
5. Stefanoff, P. *et al. Vaccine* **28,** 5731–5737 (2010).
6. Gangarosa, E. J. *et al. Lancet* **351,** 356–361 (1998).
7. Nicoll, A., Elliman, D. & Ross, E. *Br. Med. J.* **316,** 715–716 (1998).
8. Swansea Research Unit of the Royal College of General Practitioners. *Br. Med. J.* **282,** 23–26 (1981).
9. Petrovic, M., Roberts, R. & Ramsay, M. S. *Br. Med. J.* **322,** 82–85 (2001).
10. Leach, M. & Fairhead, J. *Vaccine Anxieties: Global Science, Child Health and Society* (Earthscan, 2007).

BILL & MELINDA GATES FOUNDATION

A child in Kano, Nigeria, receiving polio vaccine in June 2010.

Lessons from polio eradication

Ridding the world of polio requires a global initiative that tailors strategies to communities, say **Heidi J. Larson** and **Isaac Ghinai**.

Ten years ago all seemed to be going well with poliomyelitis eradication. The number of polio cases globally had dropped by 99% from an estimated 350,000 in 1988 to fewer than 500 in 2001, thanks to the Global Polio Eradication Initiative (GPEI).

But getting rid of the last 1% of cases over the past decade has been a roller-coaster ride including ridding whole nations of the disease and flare-ups in previously polio-free countries (see 'The disease that won't die easily'). Arrayed against the effort have been: logistical barriers, especially in conflict areas; management challenges; uncertain funding; waning political will; persisting anti-vaccine rumours and resistance; silent infections — healthy carriers who spread disease; and rare cases of vaccine-induced polio.

Against these odds, polio eradication has pushed on stubbornly; perhaps too stubbornly, at times, alienating some local populations by seeming overly top–down in its approach. But, the world cannot give up the fight to wipe out the disease that was paralysing 1,000 children a day 25 years ago and whose eradication is estimated to benefit the world by US$40 billion–50 billion between 1988 and 2035[1]. The alternatives are more costly — long-term measures to keep the number of cases low or risk widespread resurgence of a disabling and fatal disease[2].

Happily, much has been learned, and is still being learned, from the polio eradication initiative; in particular, why some children remain unvaccinated. Prompted by several years of fieldwork (by H.J.L.) with the United Nations on community acceptance of vaccines, our research team at the London School of Hygiene and Tropical Medicine has established an early-warning system to detect and investigate vaccine rumours and public concerns before they erupt into widespread vaccine refusals (go.nature.com/zfvi9s).

Our research points to three key lessons for the endgame of polio eradication and for other immunization initiatives in the developing world. First, integrate social and political analyses into feasibility assessments, strategic planning and steering. Second, find out what is driving rumours and resistance. And third, design and monitor communication and engagement strategies that work hand in hand with technical strategies and enable local populations to feel ownership of their immunization programme[3].

THE PROBLEMS

To explore how rumours can snowball into a crisis, events in Nigeria and India are worth a closer look.

What happened in Nigeria in 2003 has become a case study in the importance of getting local populations on side early[4]. Five states in the predominantly Muslim north of Nigeria — Kano, Zamfara, Kaduna, Niger and Bauchi — boycotted polio vaccination when religious and political leaders endorsed rumours that oral polio vaccine was an American conspiracy to spread HIV and cause infertility. The rumours had circulated in Nigeria and elsewhere for many years, but the tense political situation following elections in April 2003 provided motives for state governments in the north to "make things difficult" for the federal government[5]. This happened against a background of intensifying polio-eradication campaigns in May 2003, international conflicts against Muslim countries, and court proceedings in the United States where Nigerian families were suing Pfizer for allegedy unethical proceedings during clinical trials of an antibiotic drug in Kano[4].

In most Nigerian states, the vaccine suspensions were short-lived. But the newly elected governor of Kano — the most populous state, home to about 10 million people — enforced the boycott for 11 months. This catalysed a resurgence of polio in the country, with more than five times the number of cases in 2006 than in 2002 (reported incidence jumped from 202 in 2002 to 1,143 in 2006). Nigerian strains of the virus spread to 15 other countries[6], many of which had been previously certified polio-free, and were detected as far away as Indonesia.

In India, resistance to vaccination came from within similar socio-economically

VACCINES
New promise, old doubts
nature.com/vaccines

marginalized, largely Muslim, communities that were also influenced by rumours that the polio vaccine was a Western ploy to sterilize Muslims.

A dramatic increase in polio cases — from 268 in 2001 to 1,600 in 2002 — led to an investigation. It revealed that more than 80% of the children infected in the 2002 outbreak were Muslim boys under two years old, and 80% were from the state of Uttar Pradesh — one of the poorest states in India[7].

Vaccine resistance in India varied from the overt to the covert. During house-to-house visits with UNICEF to communities in Uttar Pradesh, we were sometimes told there were no children present, only to hear a baby crying in a back room. Other families closed their windows and doors when they heard vaccinators approaching. One vaccinator showed us scratches on her arms where household members had physically resisted immunization. Although mothers were often the ones to say no to health workers, their reasons for doing so often pointed to the influence of a husband or a powerful mother-in-law.

There are similarities between the Indian and Nigerian experiences. Vaccine refusals were centred on marginalized communities that lacked other basic services, such as clean water, and were suspicious of frequent door-to-door, free, polio vaccinations. Both settings involved communities responding to perceived external threats (Western conflicts or minority status) and in both, vaccine refusers became acceptors through public engagement.

THE SOLUTIONS

Faced by a sequence of such crises, the GPEI recognized that it needed a new way of working. Didactic, mass-communication approaches — such as street banners, posters and radio announcements — were doing little to persuade the most marginalized and the resistant populations.

In India, in response to the 2002 outbreak, the GPEI developed an ambitious strategy, working more closely with formal and informal social networks[8] and through local institutions such as the Aligarh Muslim University in Uttar Pradesh and the National Islamic University in New Delhi[7]. Community members were trained and deployed as mobilizers and became local 'champions' for polio eradication, countering resistance to vaccination from within their communities. The significant decline in polio cases in Uttar Pradesh is testament to the success of these relationships. The state has not seen a case of polio for more than a year.

There was also a realization that the effectiveness of engagement strategies needed to be measured by the outcomes, namely the number of children vaccinated and the number of polio cases — not just the number of community meetings or posters promoting vaccination and announcing immunization days[3].

Another innovation is the mapping of key influencers of vaccine acceptance or refusal. In Kano, Nigeria, for example, each mosque, market, school and household is plotted on a map and visited by vaccinators. Understanding the role of local traditional, religious and political leaders in India and Nigeria was essential. The visit of American philanthropist Bill Gates in early 2009, and his personal advocacy with the sultan of Sokoto and with the governor of Kano, were crucial in renewing the commitment of the states in northern Nigeria to eradicate polio.

Polio remains endemic in Afghanistan, India, Nigeria and Pakistan. There were 1,351 cases globally last year. The good news reported by the GPEI's Independent Monitoring Board[2] is that polio cases fell by more than 90% in India and Nigeria in 2010. And, even in Afghanistan, the numbers dropped by 34%. The most worrying news is that the number of polio cases in Pakistan increased by 62% in 2010. This is because of a convergence of waning political will and competing priorities such as the catastrophic floods in 2010, persisting vaccine rumours and refusals, and health-worker fatigue[9]. Mobilizing political will and engaging the public will be crucial both to Pakistan's success and to the global effort.

In 2009, the GPEI commissioned country-specific evaluations on the major barriers to polio eradication. What they revealed is how locally varied the barriers are and that "there is no single 'right way' to engage with communities"[10]. In some situations, highly visible involvement of political leaders has promoted polio vaccination — the governor of Kano vaccinating his own child after the boycott, for example. In Afghanistan, however, the evaluation recommended that "the visible involvement of political figures in vaccination campaigns" be reduced to reflect the political neutrality of the programme in a politically sensitive environment.

"Uttar Pradesh has not seen a case of polio for more than a year."

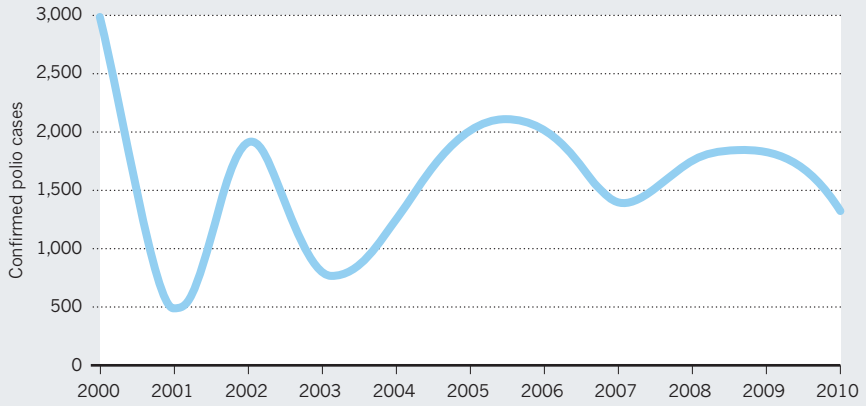
For the last leg of the race to eradicate polio, health workers must engage marginalized communities and listen to local concerns before pushing ahead with a strategy. The same lessons should be applied to other vaccination campaigns. And we must all recognize that humans are as challenging, if not more so, than the virus itself. ■

Heidi J. Larson *and* **Isaac Ghinai** *are in the Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1 7HT, UK.*
e-mail: heidi.larson@lshtm.ac.uk

1. Duintjer Tebbens, R. J. *et al. Vaccine* **29,** 334–343 (2010).
2. Donaldson, L. *et al. Report of the Independent Monitoring Board of the Global Polio Eradication Initiative* (2011); available at http://go.nature.com/emsvrv
3. Taylor, S. & Shimp, L. *J. Health Commun.* **15** (suppl. 1), 48–65 (2010).
4. Yahya, M. *Polio Vaccines — Difficult to Swallow: The Story of a Controversy in Northern Nigeria.* IDS Working Paper 261 (2006); available at http://go.nature.com/wkwomf
5. Kaufmann, J. R. & Feldbaum, H. *Health Aff.* **28,** 1091–1101 (2009).
6. Centres for Disease Control and Prevention *Morb. Mortal. Wkly Rep.* **58,** 357–362 (2009).
7. UNICEF Regional Office for South Asia. *When Every Child Counts — Engaging the Underserved Communities for Polio Eradication in Uttar Pradesh, India.* (2004); available at http://go.nature.com/t9xkl5
8. Obregón, R. *et al. Bull. World Health Organ.* **87,** 624–630 (2009).
9. Closser, S. *Chasing Polio in Pakistan: Why the World's Largest Public Health Initiative May Fail* (Vanderbilt Univ. Press, 2010).
10. Tool, M., Simmonds, S., Coghlan, B. & Mojadidi, N. *Evaluation of the Global Polio Eradication Initiative: Report on the Independent Evaluation of the Major Barriers to Interrupting Poliovirus Transmission in Afghanistan* (2009); available at http://go.nature.com/vosmoo

THE DISEASE THAT WON'T DIE EASILY

Global polio cases numbered in the hundreds of thousands in the 1980s and early 1990s. Eradication efforts reduced them to as few as 500 in 2001, but the disease lingers on.



Vast reserves of lithium carbonate salts in Bolivia and elsewhere in South America could sustain vehicle battery manufacture for centuries.

PHOTOLIBRARY.COM

TECHNOLOGY

Charging towards the superbattery

Lithium-ion technology is bringing us closer to solving energy and transport problems, finds **Bruno Scrosati**.

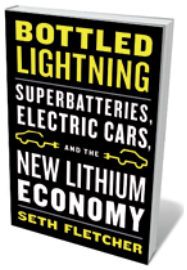
When, in 1801, Alessandro Volta unveiled his 'electric pile' gadget to Napoleon Bonaparte, he could not have imagined that, two centuries later, his invention would be central to human life. His primitive electrical cell of zinc and silver electrodes separated by a brine-soaked felt led to the compact electrochemical power source that dominates modern consumer electronics — the lithium battery.

In *Bottled Lightning*, science journalist Seth Fletcher explains how lithium batteries work and describes the research steps that have led to their ubiquity. The mobile electronics market is booming, producing billions of units a year and billions of dollars in profits. And new challenges for lithium batteries are opening up in green energy. Fletcher describes the fierce competition to develop the next generation of lithium batteries, but could have given more people their due for the existing technology.

Decreasing oil resources and concerns about climate change necessitate greater use of alternative energy sources, such as solar and wind, and the replacement of polluting internal-combustion cars with hybrid vehicles, plug-in hybrid vehicles and, ultimately, fully electric vehicles. As the sun does not always shine and the wind does not blow on command, the success of these renewable sources depends on efficient storage. Electrochemical batteries, lithium ones in particular, are the best option, converting stored chemical energy into electricity with high efficiency and without toxic emissions.

NATURE.COM
For another review about electric cars:
go.nature.com/azklzq

As yet, lithium batteries do not meet the technical requirements of hybrid or electric vehicles. The challenge is to move beyond the present chemistry to produce batteries that are safer, cheaper and have greater energy density. This will not be easy. But the ecological, economical and political rewards are so great that many countries are directing tremendous amounts of funding towards research and development in battery technology. The result, as Fletcher puts it, is that in the past decade, "advanced-battery start-ups started popping up like mushrooms after a spring rain".



Bottled Lightning: Superbatteries, Electric Cars, and the New Lithium Economy
SETH FLETCHER
Hill & Wang: 2011. 272 pp. $26, £18.99

This intense activity has also given rise to a series of patent conflicts and legal battles over priority, which Fletcher aptly calls "lithium wars". He focuses on the many-sided battle for the patent of the lithium-battery cathode material — a lithium-iron phosphate with an olivine crystal structure that is one of the most promising advanced electrode materials. As he says, such clashes are not new: patent disputes and get-rich-quick hype have dogged the battery business since its inception.

Fortunately, the battery-science community has avoided this bad atmosphere

and continues to make progress. As well as lithium-iron phosphate, other innovative materials have been used for the three main battery components of anode, cathode and electrolyte. But there is still no lithium battery light enough to power a small electric car over a reasonable distance on a single charge.

Urgently needed are 'superbatteries' with energy densities at least two or three times higher than at present. The most promising candidates are lithium–sulphur and lithium–air batteries, which in principle should be able to store 5–10 times the energy of today's cells. These are conceptually simple, but their implementation has been stalled by a series of apparently insurmountable hurdles: the high solubility of the (polysulphide) discharge products; the high resistance of the electrode materials in the case of lithium sulphur; the slow kinetics of the oxygen electrode; and the instability of the lithium anode in the case of lithium air.

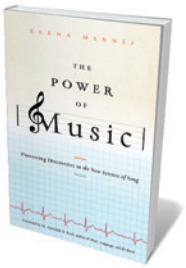
There have been breakthroughs in the past few years with the development of advanced sulphur electrode nanomorphologies, the clarification of the oxygen reduction process, the use of appropriate catalysts for promoting its evolution, and the stabilization of the lithium electrode by covering it with protective films. The road to applications is still long, but the race for the electric car has started. Many car makers are seeking joint ventures with battery manufacturers to pursue the Japanese frontrunners who, having won their early bet on hybrids, are still the major players in electric vehicles.

With demand for lithium set to grow, some question whether Earth's crust contains enough of the metal to sustain its use in vehicles. Fletcher cleverly analyses the debate and gives vivid descriptions of his trips to Bolivia and Chile to visit the two main salt deposits that, together with a third in Argentina, are the richest sources of lithium carbonate. The reserves could last for centuries, so there will be enough lithium to fill up our tanks even in the improbable case of all cars becoming hybrid or electric.

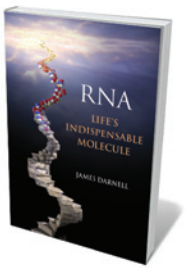
Bottled Lightning is a gripping introduction to this sophisticated technology and its place in our society. My only criticism is that Fletcher fails to credit the group of US and European scientists, including Don W. Murphy, Michel Armand and myself, who in the early 1980s developed the lithium-ion battery concept. The field then fell silent for more than ten years, until the Japanese company Sony optimized the idea for the first commercial lithium-ion battery in the early 1990s. As Fletcher notes, plenty has happened since. ■

Bruno Scrosati *is senior professor of electrochemistry at the University of Rome Sapienza, Italy.*
e-mail: bruno.scrosati@uniroma1.it

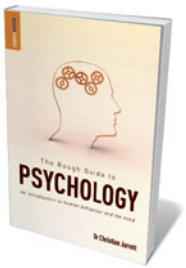
Books in brief



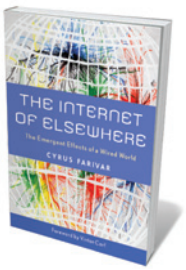
The Power of Music: Pioneering Discoveries in the New Science of Song
Elena Mannes WALKER *288 pp. $26 (2011)*
Why does music move us? In a wide-ranging book that spans science and culture, documentary-maker Elena Mannes — who hails from a long line of musicians and patrons, including the builder of New York's Carnegie Hall — describes what the latest cognitive biology and neuroscience tell us about our emotional responses to music. She points to evidence that music can heal, and looks at why music seems to be almost universal across different cultures.



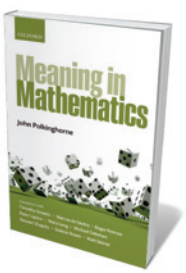
RNA: Life's Indispensable Molecule
James Darnell COLD SPRING HARBOR LABORATORY PRESS *416 pp. $39 (2011)*
The RNA molecule is crucial for gene expression and protein synthesis. Molecular biologist and RNA expert James Darnell rounds up the latest findings on RNA research in this book aimed at biology graduates. He describes how RNA's varied biochemical and structural properties were discovered, how messenger RNAs are generated and produce proteins, how RNA molecules take on regulatory roles in the cell, and how RNAs might have initiated life on Earth.



The Rough Guide to Psychology: An Introduction to Human Behaviour and the Mind
Christian Jarret ROUGH GUIDES *376 pp. £11.99 (2011)*
The basics of psychology are outlined in the latest title in the Rough Guide science series. Starting from an individual perspective, journalist Christian Jarret explores the mind and the brain, touching on memory, intelligence and personality. He goes on to analyse our relationships with others, including how we choose our friends and partners. He covers the psychological basis of crime, learning, sport, politics and shopping, as well as conditions of impaired mental health such as depression, anxiety and schizophrenia.



The Internet of Elsewhere: The Emergent Effects of a Wired World
Cyrus Farivar RUTGERS UNIVERSITY PRESS *296 pp. $25.95 (2011)*
Much of the power of the Internet — good and bad — stems from its global reach. Technology journalist and broadcaster Cyrus Farivar profiles web pioneers in four countries — Iran, Estonia, South Korea and Senegal — to illustrate how the Internet is transforming international communications, politics and economics. His case studies examine the Internet's history and effects in these diverse nations, showing that they are at the forefront of developments in Internet phone services, broadband access and digital law.



Meaning in Mathematics
Edited by John Polkinghorne OXFORD UNIVERSITY PRESS *192 pp. £18.99 (2011)*
Is mathematics discovered or invented? Nine top scholars, including mathematical physicist Roger Penrose and philosopher Gideon Rosen, muse on whether the discipline is a purely intellectual pursuit or a means of uncovering real aspects of nature. Intended for a broad audience, each essay in this volume — edited by mathematician-turned-theologian John Polkinghorne — is accompanied by comments from the other contributors.

and continues to make progress. As well as lithium-iron phosphate, other innovative materials have been used for the three main battery components of anode, cathode and electrolyte. But there is still no lithium battery light enough to power a small electric car over a reasonable distance on a single charge.

Urgently needed are 'superbatteries' with energy densities at least two or three times higher than at present. The most promising candidates are lithium–sulphur and lithium–air batteries, which in principle should be able to store 5–10 times the energy of today's cells. These are conceptually simple, but their implementation has been stalled by a series of apparently insurmountable hurdles: the high solubility of the (polysulphide) discharge products; the high resistance of the electrode materials in the case of lithium sulphur; the slow kinetics of the oxygen electrode; and the instability of the lithium anode in the case of lithium air.

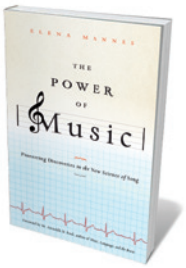
There have been breakthroughs in the past few years with the development of advanced sulphur electrode nanomorphologies, the clarification of the oxygen reduction process, the use of appropriate catalysts for promoting its evolution, and the stabilization of the lithium electrode by covering it with protective films. The road to applications is still long, but the race for the electric car has started. Many car makers are seeking joint ventures with battery manufacturers to pursue the Japanese frontrunners who, having won their early bet on hybrids, are still the major players in electric vehicles.

With demand for lithium set to grow, some question whether Earth's crust contains enough of the metal to sustain its use in vehicles. Fletcher cleverly analyses the debate and gives vivid descriptions of his trips to Bolivia and Chile to visit the two main salt deposits that, together with a third in Argentina, are the richest sources of lithium carbonate. The reserves could last for centuries, so there will be enough lithium to fill up our tanks even in the improbable case of all cars becoming hybrid or electric.

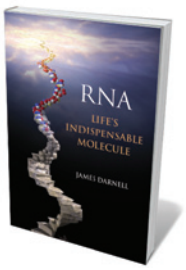
Bottled Lightning is a gripping introduction to this sophisticated technology and its place in our society. My only criticism is that Fletcher fails to credit the group of US and European scientists, including Don W. Murphy, Michel Armand and myself, who in the early 1980s developed the lithium-ion battery concept. The field then fell silent for more than ten years, until the Japanese company Sony optimized the idea for the first commercial lithium-ion battery in the early 1990s. As Fletcher notes, plenty has happened since. ■

Bruno Scrosati *is senior professor of electrochemistry at the University of Rome Sapienza, Italy.*
e-mail: bruno.scrosati@uniroma1.it

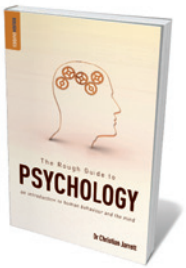
Books in brief



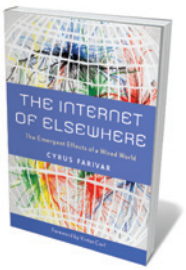
The Power of Music: Pioneering Discoveries in the New Science of Song
Elena Mannes WALKER *288 pp. $26 (2011)*
Why does music move us? In a wide-ranging book that spans science and culture, documentary-maker Elena Mannes — who hails from a long line of musicians and patrons, including the builder of New York's Carnegie Hall — describes what the latest cognitive biology and neuroscience tell us about our emotional responses to music. She points to evidence that music can heal, and looks at why music seems to be almost universal across different cultures.



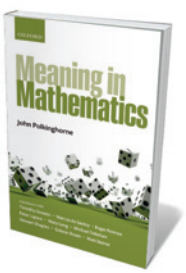
RNA: Life's Indispensable Molecule
James Darnell COLD SPRING HARBOR LABORATORY PRESS *416 pp. $39 (2011)*
The RNA molecule is crucial for gene expression and protein synthesis. Molecular biologist and RNA expert James Darnell rounds up the latest findings on RNA research in this book aimed at biology graduates. He describes how RNA's varied biochemical and structural properties were discovered, how messenger RNAs are generated and produce proteins, how RNA molecules take on regulatory roles in the cell, and how RNAs might have initiated life on Earth.



The Rough Guide to Psychology: An Introduction to Human Behaviour and the Mind
Christian Jarret ROUGH GUIDES *376 pp. £11.99 (2011)*
The basics of psychology are outlined in the latest title in the Rough Guide science series. Starting from an individual perspective, journalist Christian Jarret explores the mind and the brain, touching on memory, intelligence and personality. He goes on to analyse our relationships with others, including how we choose our friends and partners. He covers the psychological basis of crime, learning, sport, politics and shopping, as well as conditions of impaired mental health such as depression, anxiety and schizophrenia.



The Internet of Elsewhere: The Emergent Effects of a Wired World
Cyrus Farivar RUTGERS UNIVERSITY PRESS *296 pp. $25.95 (2011)*
Much of the power of the Internet — good and bad — stems from its global reach. Technology journalist and broadcaster Cyrus Farivar profiles web pioneers in four countries — Iran, Estonia, South Korea and Senegal — to illustrate how the Internet is transforming international communications, politics and economics. His case studies examine the Internet's history and effects in these diverse nations, showing that they are at the forefront of developments in Internet phone services, broadband access and digital law.



Meaning in Mathematics
Edited by John Polkinghorne OXFORD UNIVERSITY PRESS *192 pp. £18.99 (2011)*
Is mathematics discovered or invented? Nine top scholars, including mathematical physicist Roger Penrose and philosopher Gideon Rosen, muse on whether the discipline is a purely intellectual pursuit or a means of uncovering real aspects of nature. Intended for a broad audience, each essay in this volume — edited by mathematician-turned-theologian John Polkinghorne — is accompanied by comments from the other contributors.

S. HARRIS

NEUROSCIENCE

What makes us laugh

Humour is the brain's reward for discovering unexpected errors, says **Appletree Rodden**.

Photons have mass? I didn't even know they were Catholic.

Why do some of us find that funny? *Inside Jokes* surveys the scientific basis of humour and proposes a new theory. It presents a brief history of the concept's development from the ancient Greeks to the present, discusses the possible origin of laughter from a Darwinian perspective and describes what is known about jocularity in the brain.

Co-authored with philosopher Daniel Dennett and psychologist Reginald Adams, the book grew out of the dissertation of neuroscientist Matthew Hurley, then at Tufts University in Medford, Massachusetts. The authors' account of why humour and laughter exist independently, and how they relate — such that laughter sometimes "expresses the detection of humour" — is a valuable, if not a full, explanation. A mix of lightness and seriousness, the book also contains a great collection of jokes: from awful groaners to choice quips.

The authors propose that humour is a cognitive event, in which an unconscious assumption is discovered to have been a mistake. For example, if on reading this magazine you suddenly become aware that the pages are not made of paper but of pressed haddock, then, they argue, that should strike you as humorous. The countering of the belief that 'the journal *Nature* is never printed on pressed haddock' should not merely surprise you, but seem funny. However, danger trumps humour: it would not be amusing, for example, if the pressed haddock were radioactive.

"Countering the belief that 'the journal Nature is never printed on pressed haddock' should seem funny."

Inside Jokes: Using Humor to Reverse-Engineer the Mind
MATTHEW M. HURLEY, DANIEL C. DENNETT AND REGINALD B. ADAMS JR
MIT Press: 2011. 384 pp. $29.95, £22.95

Humour, the authors suggest, is an element in the cognitive 'just-in-time spreading activation' system, by which our brains fit the best overall meaning to the collection of mental scripts or frames it has at its disposal. For example, you may currently be following a 'reading the journal *Nature*' script; perhaps also a 'sitting at my desk' script, and maybe a 'trying not to forget to stop at the grocery store on the way home' script. The brain is constantly mediating among these frames, charting our course through a fast-moving, life-threatening world.

Our ability to fashion 'just in time' meaning from this jumble is far from perfect. We are not computers but flesh-and-blood, jerry-built synthesizers of meaning constructed from the impressions provided by our sense organs, memories and emotions. Humour happens when this operating system detects an error that other parts had overlooked. The brain's dopaminergic pleasure system rewards that survival-benefiting discovery with a jolt of mirth.

The authors explain how their ideas build on previous theories of how humour emerges. Notably, that it comes from the joke-teller's position of superiority, as proposed by Aristotle and Thomas Hobbes; when an incongruity is resolved, as suggested by Immanuel Kant, Arthur Schopenhauer and V. S. Ramachandran; on release from internal censors, following Sigmund Freud; and from some kinds of surprise, as hypothesized by psychologist Jerry M. Suls. It can also be inspired by shifting our frame of reference, according to Marvin Minsky, Victor Raskin and Salvatore Attardo.

The authors sometimes labour to bend the phenomena of humour and laughter to fit their theory. And their attempt to explain every reason why humans laugh, smile or experience low-grade mirth is not entirely satisfying. To their credit, the authors realize this, and rightly consider their book a valuable contribution. ■

Appletree Rodden *is a biochemist, physician and cognitive scientist at the Christian Hospital of Quakenbrueck, Germany.*
e-mail: annetree@aol.com

P. BOLGER/SCIENCE GALLERY, TCD

Euthanasia Coaster **by Julijonas Urbonas imagines a thrilling way out should we become bored of artificially extended lifespans: crushing death by roller coaster.**

ART

Body work

Genetics and artificial intelligence figure prominently in an unsettling Dublin exhibition, discovers **Anthony King**.

We all wonder about tomorrow. The *Human+* exhibition at Dublin's Science Gallery speculates on how science and technology might enhance our bodies and well-being and transform what it means to be human. Genetics and artificial intelligence figure prominently among its themes of augmented abilities, authoring evolution, extended ecologies, life at the edges and non-human encounters.

The exhibition, supported by the Wellcome Trust, was developed with the Trinity Long Room Hub, the university's new centre for humanities, and Trinity College Dublin's School of Medicine, which is celebrating its tercentenary.

People traditionally viewed as disabled are often early adopters of new enhancement technologies, says Michael John Gorman, director of the Science Gallery. The photographic portraits of Aimee Mullins challenge our notions of beauty and athleticism. As a child, both her legs were amputated below her knee, so she learned to walk on prosthetic legs and later set world records running on carbon-fibre prostheses.

Others choose to make radical upgrades to their body. Nina Sellars's *Oblique* displays photographs of the performance artist Stelarc under the surgeon's knife as he begins the process of implanting an ear on his arm. Stelarc's *Prosthetic Head*, a giant on-screen avatar whom visitors can interrogate through a keyboard, has been a major draw. The head, occupying a room of its own, displays real-time lip synching, speech synthesis and facial expressions — and attitude. Abuse is met with threats: "I will remember you said that when robots take over the world."

Human+ focuses more on human–robot interactions than cyborg mechanics or gadgetry. A wall of robotic eyes tracks you eerily through *Area V5* by Louis-Philippe Demers. The sense of unease grows with the seemingly more human *My Robot Companion*, a freakish white plastic visage that mimics the facial features of those around it. Scattered electronic parts in Yves Gellie's photographs of robotics labs suggest the imminent creation of life. But it is unclear whether the nano-sized machines boring holes into the human body in the film *Aphasia Mechanica* are busy repairing or destroying their host.

P. BOLGER/SCIENCE GALLERY, TCD

Detail from *Area V5*: moving eyes trigger unease.

Human+: The Future of Our Species
Science Gallery, Trinity College Dublin. Until 24 June 2011.

Looming above the gallery staircase is the wax colossus *If Not Now Then When*. The hooded and bloated figure, in a classical pose, forces us to confront the effects of our over-consumption, waste and pollution. As in Oscar Wilde's 1890 novel *The Picture of Dorian Gray*, in which the painting ages and absorbs the evil deeds of its sitter, the wax figure can be seen as a kind of sponge, says its creator, John Isaacs.

Taking a still darker turn is the sculpture *Euthanasia Coaster*. Should medical wonders allow us extended lifetimes, boredom may bedevil us. Julijonas Urbonas imagines a humane and thrilling exit: death by roller coaster in the form of an exhilarating 500-metre drop followed by a series of loops, the *G*-forces of which would kill passengers in a state of intense euphoria.

Science students from the university are on hand to discuss each work. "The conversation with the public can get quite deep," says geneticist Aoife McLysaght. And visitors can participate: DNA samples taken from around 200 visitors each week will be tested for the dopamine D4 receptor gene, a variant of which has been linked to high-risk behaviour.

Human+ is not all fizzes and bangs: its thoughtful works provoke visitors into deeper engagement with important societal issues. ■

NATURE.COM
For a review of Phil Ball's book on making people:
go.nature.com/jgmb15

Anthony King *is a writer based in Dublin.*
e-mail: anthonyjking@gmail.com

CORRESPONDENCE

Peer reviews: in praise of referees

Unlike Hidde Ploegh, I am grateful to reviewers who suggest lateral experiments (*Nature* **472,** 391; 2011). Good science depends on reproducible results, and the reviewers are often just calling on authors to replicate their results by different means.

Ploegh is critical of the cost and extra time needed to do more experiments, but what about the cost in wasted time when published results cannot be replicated? In my experience, the lateral experiments are usually better than those the authors planned to do next anyway. They often strengthen the original results and lead to useful discoveries.

Reviewers are doing authors a great favour in suggesting specific, focused experiments; they subsequently spend (unpaid) time re-reviewing the paper. Rather than criticism, they deserve a resounding thanks.

Eric L. Altschuler *New Jersey Medical School, New Jersey, USA. altschel@umdnj.edu*

Peer reviews: make them public

Making peer reviewers' comments public — not necessarily signed — would alleviate most of the problems outlined by Hidde Ploegh (*Nature* **472,** 391; 2011).

Readers of the comments would then be able to judge, for example, whether reviewer requests for additional experiments were reasonable. Such a public-review policy would help editors and add a new dimension to a journal's reputation, particularly if others in the field publicly shared their own relevant observations.

In conventional peer review, especially at top-tier journals, much of the reviewing effort goes into manuscripts that are ultimately rejected, meaning that the scientific community has no access to these communications. Under a public system, these records could prevent reinventions of the wheel and help educate newcomers to the field or to peer reviewing.

Publicly available reviews, including those of rejected manuscripts, would also provide an incentive for authors to submit their manuscript only when it is ready — helping to lower rejection rates and aiding the search for suitable reviewers (see go.nature.com/qamrfc).

Daniel Mietchen *EvoMRI Consulting, Jena, Germany. daniel.mietchen@evomri.net*

Cooperation is key to Asian hydropower

Environmentalists won a reprieve last month against construction of the Xayaburi dam on the lower Mekong River in Laos (*Nature* doi:10.1038/news.2011.220). But it is the Laos government that will have the final say.

China is leading this hydropower boom in southeast Asia, and aims to increase its hydropower from 200 to 380 gigawatts by 2020. Dams are planned or completed at sites along other international rivers, including 13 on the Salween or Nujiang, which is protected by UNESCO (United Nations Educational, Scientific and Cultural Organization), and 20 along the Brahmaputra — all in rare and fragile environments.

In 2004, Chinese Premier Wen Jiabao halted the development of dams along the Salween. But the order was lifted after the National Development and Reform Commission called last year for dam building to proceed, prompted by soaring power demands and the energy and water conservation targets of China's latest five-year plan.

This wave of development in hydropower and its effect on water resources is likely to intensify water-related disputes among neighbouring riparian countries. To assess properly the impact of building hydropower dams, transparent policies and multinational cooperation are crucial.

Lishan Ran, X. X. Lu *National University of Singapore, Singapore. geoluxx@nus.edu.sg*

Seeking out Earth's warning signals

I disagree with Robert Geller's hard-line stance against earthquake prediction (*Nature* **472,** 407–409; 2011). Although early warning signs are diverse, fleeting and often subtle, they can also be surprisingly strong, even for moderate earthquakes (see, for example, T. Bleier *et al. Nat. Hazards Earth Syst. Sci.* **9,** 585–603; 2009).

More than 100 years of seismology have led to an advanced understanding of the tectonic forces that cause Earth's plates to move, slide past each other and collide. But when it comes to earthquake prediction, the seismological approach has always been to try to understand how past events happened and to develop probability models for 'predicting' when the next ones might occur. This analysis has built-in statistical uncertainties that are of the order of years, decades, even centuries — and there is no way around it.

Any good seismologist will recognize the limitations of earthquake prediction. But the study of earthquakes should include the tracking down and investigation of all the different signals that Earth produces before a catastrophic rupture. If seismologists can't do it alone, can't we do it collectively across disciplines?

Friedemann Freund *NASA Ames Research Center, SETI Institute and San Jose State University, California, USA. friedemann.t.freund@nasa.gov*

Can Facebook influence funding?

I would like to make it clear that I played no part in instigating a Facebook uprising over my research (*Nature* **472,** 410–411; 2011).

I am not an activist but a scientist who has published 27 peer-reviewed studies of chronic cerebrospinal venous insufficiency (CCSVI) and its relationship to multiple sclerosis in 18 interdisciplinary journals.

This research was funded by the Italian government and banking foundations, and grants were peer-reviewed by scientific committees under the usual rules.

I do not believe that Facebook can influence the diversion of funds to change research priorities or the judgement of the scientific community.

CCSVI is a pathological condition first described in the literature two years ago. A Google Scholar search reveals that CCSVI has been cited more than 2,000 times in published scientific papers. Evidently, CCSVI is a hot topic — it is interesting precisely because it is controversial.

Paolo Zamboni *University of Ferrara, Italy. zambo@unife.it*
Competing financial interests declared (see http://dx.doi.org/10.1038/473452e).

CONTRIBUTIONS

Correspondence may be submitted to **correspondence@nature.com** after consulting the author guidelines at **http://go.nature.com/cmchno**.

ENTOMOLOGY

Royal aspirations

What makes a queen honeybee? The proposal of a definitive answer to this long-standing question offers much royal food for thought for those studying the evolution of social traits and insect genomes. SEE ARTICLE P.478

GENE E. ROBINSON

A queen honeybee (*Apis mellifera*) has 100 times greater reproductive capacity and lives 20 times longer than the other female denizens of the beehive, the workers. These differences are due to royal jelly, which is composed mainly of water, protein, sugars, lipids and minerals, and which is fed copiously to young larvae when a bee colony needs a new queen. Noting these differences, hopeful humans have made royal jelly a popular dietary supplement, especially in Asia, with annual global sales of more than US$600 million[1].

But despite intense scientific and economic interest, the specific substances in royal jelly that cause this remarkable transformation had, up to now, escaped detection. An article by Kamakura on page 478 of this issue[2] finally ends this almost 100-year quest[3]. Creative coupling of honeybees and fruitflies (*Drosophila melanogaster*) has led to the discovery of the central transformative role of a protein called royalactin and of its mode of action.

The first break came when Kamakura[2] observed temperature-related differences in the potency of royal jelly, with gradual degradation at 40 °C and a complete loss of queen-inducing activity after 30 days. On screening various constituents of royal jelly stored at this temperature, he found that a protein with a molecular mass of 57,000 fitted the profile perfectly; it degraded gradually and was gone completely after 30 days. When purified and tested on larvae in a laboratory assay, the protein, royalactin, caused three classic queen-related changes: shortened developmental time, increased adult mass and larger ovaries (Fig. 1).

The second break came when Kamakura made the shocking discovery that royalactin has the same three effects on *Drosophila*. Shocking? Yes, because there are no queen flies. *Drosophila* and *Apis* are in different insect orders, separated by more than 300 million years of evolution. Moreover, the social life of fruitflies is much more limited than that of honeybees, which form one of the most complex societies on Earth.

Nevertheless, the parallel effects enabled powerful *Drosophila* genetic techniques to be used to comprehensively elucidate royalactin's mode of action. Kamakura used RNA interference (RNAi) and the Gal4 system for tissue-specific control of gene expression to show that royalactin stimulated epidermal growth factor receptor (EGFR)-mediated signalling in the fat bodies, a peripheral nutrient-sensing tissue analogous to vertebrate liver and adipose tissues. EGFR in turn activated various kinases including S6K, known in *Drosophila* to affect cell size and adult lifespan, and MAPK, which affects the duration of development, among other things. The same results were obtained with ectopic expression of royalactin in the fat bodies. Closing the loop, *Egfr* RNAi was used to confirm that royalactin acts through EGFR signalling in honeybees.

This paper[2] harks back to the earliest research on royal jelly. Biochemical analyses[4] (including a prescient study[5] using *Drosophila* back in 1948) repeatedly found promise in various royal-jelly extracts, only to wind up with no caste-determination effects. Then came endocrine analyses that not only implicated juvenile hormone in caste determination, but also diminished interest in the search for transformational substances by positing that royal jelly was nothing more than a phagostimulant[6,7]. The idea was that royal jelly's high sugar content causes the honeybee larvae to eat more of it, grow more quickly, and produce more juvenile hormone during the early larval stages, when developmental fate is sealed. Experiments showed that laboratory-reared larvae given only a meagre, worker-inducing diet could still develop into queens if also treated with juvenile hormone, or if the diet were supplemented with extra sugar.

Molecular analyses 25 years later revealed massive changes in gene expression associated with caste determination[8] that are indeed hormonally orchestrated, both by juvenile hormone and by insulin-related signalling pathways[9]. More recently, epigenetics entered the picture, thanks to the genome-sequencing-related discovery of the honeybee as the first insect known to have a fully functioning DNA-methylation system[10]. RNAi knockdown of *DNA methyltransferase 3*, a key methylation gene, increases queen development[11].

PHOTOLIBRARY.COM

Figure 1 | Centre of attention. Honeybee workers fuss over their queen — who, thanks to a larval diet of royal jelly with royalactin as the transforming agent[2], has developed faster and become bigger, with larger ovaries, than the workers have.

It seemed that a complete explanation of caste determination could eventually be fashioned from these phagostimulant, endocrine and epigenetic discoveries.

The first indication that royal jelly might not just be junk food came in a report[12] showing that (*E*)-10-hydroxy-2-decenoic acid, a major fatty-acid constituent of royal jelly, has histone deacetylase inhibitor activity. This means that a specific component of royal jelly can promote queen development by opening chromatin to permit gene activation. Add to this the new royalactin findings[2] and it is clear that royal jelly contains potent caste-determining substances after all. The challenge now is to integrate the roles in caste determination played by royalactin and the previously discovered nutritional, endocrine and epigenetic mechanisms. This might require the study of a fuller suite of queen traits, both morphological and behavioural, than explored so far.

In addition to reporting a breakthrough in one of the more productive lines of research in sociogenomics, Kamakura's paper[2] provides two general lessons, one strategic and the other conceptual. Although many researchers who study other organisms already rely bioinformatically on *Drosophila* to interpret gene function, Kamakura shows that fruitflies can also be used experimentally for this purpose, even for social traits that they lack.

Or do they? This paper provides a strong reminder that deep molecular conservation lurks beneath even the most evolutionarily novel traits. Caste determination based on royal jelly is considered to be a unique feature of honeybee social evolution, even among social insects, and no other species feed such a substance to their young. Royalactin is the product of one of ten genes that encode the major proteins in royal jelly, prime examples of novel genes[13]. These genes are probably derived from those that encode the yellow protein family, which curiously is found only in bacteria, fungi and insects. But given their involvement in caste determination, it was assumed that the royal-jelly genes became untethered from their ancestral functions and took on new tasks. They certainly have, but Kamakura's results suggest that these royal-jelly genes may not have broken all ties with the past (although it is not known whether yellow proteins in *Drosophila* also act on EGFR signalling).

We can take heart that even fruitflies harbour royal aspirations. This lesson of deep molecular conservation no doubt applies to all complex social traits, in both invertebrates and vertebrates. Long live the queen! ■

Gene E. Robinson *is at the Institute for Genomic Biology, Department of Entomology, and the Neuroscience Program, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.*
e-mail: generobi@illinois.edu

1. Euromonitor Passport Global Market Information Database (13 Nov. 2009); www.portal.euromonitor.com
2. Kamakura, M. *Nature* **473,** 478–483 (2011).
3. Aeppler, C. W. *Gleanings Bee Cult.* **50,** 151–153 (1922).
4. Wilson, E. O. *The Insect Societies* (Belknap, 1971).
5. Gardner, T. S. *J. Gerontol.* **3,** 1–8 (1948).
6. Wirtz, P. & Beetsma, J. *Entomol. Exp. Appl.* **15,** 517–520 (1972).
7. Asencot, M. & Lensky, Y. *Life Sci.* **18,** 693–699 (1976).
8. Evans, J. D. & Wheeler, D. E. *Genome Biol.* **2,** research0001.1–0001.6 (2000).
9. Patel, A. *et al. PLoS One* **2,** e509 (2007).
10. Wang, Y. *et al. Science* **314,** 645–647 (2006).
11. Kucharski, R. *et al. Science* **319,** 1827–1830 (2008).
12. Spannhoff, A. *et al. EMBO Rep.* **12,** 238–243 (2011).
13. Fischman, B. J., Woodard, S. H. & Robinson, G. E. *Proc. Natl Acad. Sci. USA* **108,** 7472–7477 (2011).

APPLIED PHYSICS

A stroke of X-ray

X-rays were discovered more than 100 years ago. They have since become a staple tool for medicine and science, so researchers are continuing their efforts to find innovative ways to produce them.

STEFAN KNEIP

A few years ago, a discovery[1] was made that flabbergasted scientists and laymen alike: peeling a common adhesive tape produces X-rays bright enough to take an image resolving a human digit. Writing in *Applied Physics Letters*, Hird *et al.*[2] now describe how they have developed a prototype that promises to turn this principle of X-ray production into a simple, low-cost X-ray source.

Since their discovery in 1895, X-rays have affected many aspects of our lives, allowing us to visualize the insides of our bodies, infer the structure of DNA and test the integrity of aircraft wings. The humble tube that first produced X-rays has seen considerable development and is still widely used. But demand for — and development of — complementary sources of X-ray radiation has also abounded. This has educed some of the most sophisticated scientific apparatuses catering for cutting-edge research, and continues to foster the development of innovative X-ray sources for routine applications[1,2].

The phenomenon that underlies the production of visible light and short bursts of X-ray emission when tape is peeled[1,3] is called triboluminescence. Analogous to sonoluminescence[4], in which energy from sound waves is converted into flashes of light, triboluminescence[5] concentrates diffuse mechanical energy into light. This can happen as a result of pulling apart, ripping, scratching or stroking material.

Light is emitted when electrons are accelerated or stopped, or when they jump between energy levels. Therefore, to obtain X-ray photons with energies of tens of kiloelectronvolts (keV) — the energies required for medical applications — electrons of at least that energy must be produced. Endowing electrons with kiloelectronvolt energies is thus a common challenge for many commercial and scientific sources of X-ray radiation. This challenge is met by high-voltage equipment, which affects safety requirements, portability, usage range and the minimum size of the sources' X-ray tubes.

The demonstration[1,3] that X-rays could be produced with an object as simple as adhesive tape, and without the application of an external source of high voltage, encouraged scientists to investigate further. Hird and colleagues' prototype[2] is the outcome of one such investigation. Their device offers the prospect of building a low-technology, economical and compact X-ray apparatus for commercial engineering, and to systematically improve our understanding of the physics of triboelectric charge transfer — the phenomenon that underlies triboluminescence.

The latest prototype fits into the hand and is intriguingly simple (Fig. 1a). It consists of an actuator that repeatedly brings an epoxy surface in and out of contact with a silicone membrane. This stroke motion causes the silicone and epoxy to acquire a charge imbalance. Triboelectrification — that is, charging up due to (frictional) contact between materials — can create high electric fields, in excess of hundreds of kilovolts per centimetre[6,7]. This is high enough to ionize the surrounding air and produce a spark — similar to the static shock that can be generated by touching an object such as a doorknob.

Hird *et al.*[2] find that, when their experimental apparatus is enclosed in a moderate vacuum, X-ray radiation is generated at a rate of more than 100,000 X-ray photons per contact cycle. The radiation is produced by atomic transitions and by decelerating electrons. This results in narrow spectral lines on top of a broad spectrum. According to their calculations, the silicone–epoxy system generates up to 10^{10} charges (electrons) per square centimetre across the contact area of the device (65 mm^2). The discharge physics is not yet fully understood[1,2], but at such charge densities, as the silicone and epoxy are separated,

It seemed that a complete explanation of caste determination could eventually be fashioned from these phagostimulant, endocrine and epigenetic discoveries.

The first indication that royal jelly might not just be junk food came in a report[12] showing that (*E*)-10-hydroxy-2-decenoic acid, a major fatty-acid constituent of royal jelly, has histone deacetylase inhibitor activity. This means that a specific component of royal jelly can promote queen development by opening chromatin to permit gene activation. Add to this the new royalactin findings[2] and it is clear that royal jelly contains potent caste-determining substances after all. The challenge now is to integrate the roles in caste determination played by royalactin and the previously discovered nutritional, endocrine and epigenetic mechanisms. This might require the study of a fuller suite of queen traits, both morphological and behavioural, than explored so far.

In addition to reporting a breakthrough in one of the more productive lines of research in sociogenomics, Kamakura's paper[2] provides two general lessons, one strategic and the other conceptual. Although many researchers who study other organisms already rely bioinformatically on *Drosophila* to interpret gene function, Kamakura shows that fruitflies can also be used experimentally for this purpose, even for social traits that they lack.

Or do they? This paper provides a strong reminder that deep molecular conservation lurks beneath even the most evolutionarily novel traits. Caste determination based on royal jelly is considered to be a unique feature of honeybee social evolution, even among social insects, and no other species feed such a substance to their young. Royalactin is the product of one of ten genes that encode the major proteins in royal jelly, prime examples of novel genes[13]. These genes are probably derived from those that encode the yellow protein family, which curiously is found only in bacteria, fungi and insects. But given their involvement in caste determination, it was assumed that the royal-jelly genes became untethered from their ancestral functions and took on new tasks. They certainly have, but Kamakura's results suggest that these royal-jelly genes may not have broken all ties with the past (although it is not known whether yellow proteins in *Drosophila* also act on EGFR signalling).

We can take heart that even fruitflies harbour royal aspirations. This lesson of deep molecular conservation no doubt applies to all complex social traits, in both invertebrates and vertebrates. Long live the queen! ■

Gene E. Robinson *is at the Institute for Genomic Biology, Department of Entomology, and the Neuroscience Program, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.*
e-mail: generobi@illinois.edu

1. Euromonitor Passport Global Market Information Database (13 Nov. 2009); www.portal.euromonitor.com
2. Kamakura, M. *Nature* **473,** 478–483 (2011).
3. Aeppler, C. W. *Gleanings Bee Cult.* **50,** 151–153 (1922).
4. Wilson, E. O. *The Insect Societies* (Belknap, 1971).
5. Gardner, T. S. *J. Gerontol.* **3,** 1–8 (1948).
6. Wirtz, P. & Beetsma, J. *Entomol. Exp. Appl.* **15,** 517–520 (1972).
7. Asencot, M. & Lensky, Y. *Life Sci.* **18,** 693–699 (1976).
8. Evans, J. D. & Wheeler, D. E. *Genome Biol.* **2,** research0001.1–0001.6 (2000).
9. Patel, A. *et al. PLoS One* **2,** e509 (2007).
10. Wang, Y. *et al. Science* **314,** 645–647 (2006).
11. Kucharski, R. *et al. Science* **319,** 1827–1830 (2008).
12. Spannhoff, A. *et al. EMBO Rep.* **12,** 238–243 (2011).
13. Fischman, B. J., Woodard, S. H. & Robinson, G. E. *Proc. Natl Acad. Sci. USA* **108,** 7472–7477 (2011).

APPLIED PHYSICS

A stroke of X-ray

X-rays were discovered more than 100 years ago. They have since become a staple tool for medicine and science, so researchers are continuing their efforts to find innovative ways to produce them.

STEFAN KNEIP

A few years ago, a discovery[1] was made that flabbergasted scientists and laymen alike: peeling a common adhesive tape produces X-rays bright enough to take an image resolving a human digit. Writing in *Applied Physics Letters*, Hird *et al.*[2] now describe how they have developed a prototype that promises to turn this principle of X-ray production into a simple, low-cost X-ray source.

Since their discovery in 1895, X-rays have affected many aspects of our lives, allowing us to visualize the insides of our bodies, infer the structure of DNA and test the integrity of aircraft wings. The humble tube that first produced X-rays has seen considerable development and is still widely used. But demand for — and development of — complementary sources of X-ray radiation has also abounded. This has educed some of the most sophisticated scientific apparatuses catering for cutting-edge research, and continues to foster the development of innovative X-ray sources for routine applications[1,2].

The phenomenon that underlies the production of visible light and short bursts of X-ray emission when tape is peeled[1,3] is called triboluminescence. Analogous to sonoluminescence[4], in which energy from sound waves is converted into flashes of light, triboluminescence[5] concentrates diffuse mechanical energy into light. This can happen as a result of pulling apart, ripping, scratching or stroking material.

Light is emitted when electrons are accelerated or stopped, or when they jump between energy levels. Therefore, to obtain X-ray photons with energies of tens of kiloelectronvolts (keV) — the energies required for medical applications — electrons of at least that energy must be produced. Endowing electrons with kiloelectronvolt energies is thus a common challenge for many commercial and scientific sources of X-ray radiation. This challenge is met by high-voltage equipment, which affects safety requirements, portability, usage range and the minimum size of the sources' X-ray tubes.

The demonstration[1,3] that X-rays could be produced with an object as simple as adhesive tape, and without the application of an external source of high voltage, encouraged scientists to investigate further. Hird and colleagues' prototype[2] is the outcome of one such investigation. Their device offers the prospect of building a low-technology, economical and compact X-ray apparatus for commercial engineering, and to systematically improve our understanding of the physics of triboelectric charge transfer — the phenomenon that underlies triboluminescence.

The latest prototype fits into the hand and is intriguingly simple (Fig. 1a). It consists of an actuator that repeatedly brings an epoxy surface in and out of contact with a silicone membrane. This stroke motion causes the silicone and epoxy to acquire a charge imbalance. Triboelectrification — that is, charging up due to (frictional) contact between materials — can create high electric fields, in excess of hundreds of kilovolts per centimetre[6,7]. This is high enough to ionize the surrounding air and produce a spark — similar to the static shock that can be generated by touching an object such as a doorknob.

Hird *et al.*[2] find that, when their experimental apparatus is enclosed in a moderate vacuum, X-ray radiation is generated at a rate of more than 100,000 X-ray photons per contact cycle. The radiation is produced by atomic transitions and by decelerating electrons. This results in narrow spectral lines on top of a broad spectrum. According to their calculations, the silicone–epoxy system generates up to 10^{10} charges (electrons) per square centimetre across the contact area of the device (65 mm^2). The discharge physics is not yet fully understood[1,2], but at such charge densities, as the silicone and epoxy are separated,

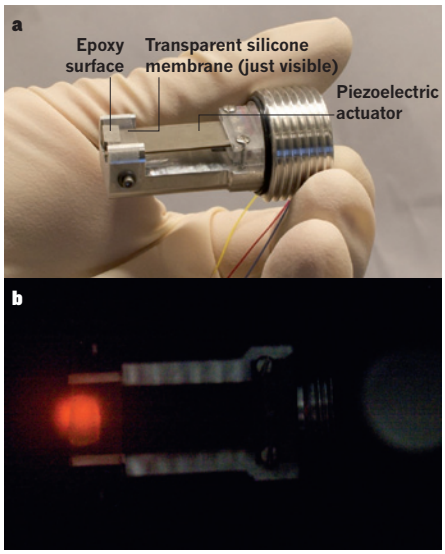


Figure 1 | A triboelectric X-ray machine. **a**, Hird and colleagues' hand-held device, shown here without its vacuum encapsulation, produces X-rays (not shown) when a piezoelectric actuator repeatedly brings an epoxy surface in and out of contact with a silicone membrane. **b**, The image shows the vacuum-enclosed apparatus in operation in a low-pressure neon atmosphere, in which the characteristic orange–red of neon glow discharge can be clearly seen. (Images courtesy of UCLA.)

the ambient gas that surrounds the vacuum-enclosed apparatus should be ionized. This gas ionization is confirmed by the characteristic orange–red of neon glow discharge that is seen when the X-ray source is operated in a low-pressure neon environment (Fig. 1b).

The authors chose silicone because of its strong tendency to charge negatively; a list known as the triboelectric series[8] exists that ranks a material according to its propensity to become charged. To test their device, they added silver to the epoxy surface, and observed the characteristic X-ray K-line emissions (around 22–25 keV) of silver. In doing so, Hird *et al.* prove conclusively that electrons from the discharge process are accelerated to tens of kiloelectronvolts of energy and produce X-rays on impact with the epoxy. The capability to load epoxy with materials of different atomic number gives the apparatus flexibility — it can both be tuned to a desired X-ray line-emission energy and increase the efficiency with which X-rays are generated.

What's more, Hird and colleagues[2] find that, by reducing the ambient pressure, the X-ray emission can persist for more than a second after the epoxy and silicone have been separated. However, to increase the X-ray yield with contact-cycle frequency, it is more desirable to achieve short bursts of X-ray emission. Although this is at the expense of X-ray energy, the emission time can be reduced to less than 10 milliseconds when the device is operated at a higher ambient pressure of 30 millitorr of nitrogen. At such pressure, the authors were able to demonstrate linear scaling of photon number with the frequency of contact cycles. This scaling suggests that the limiting factor to achieving a photon yield of 10^8 per second is finding a linear actuator capable of millimetre displacement and a contact-cycle frequency of 0.1–1 kHz. Their 'mark 1' device was based on a solenoid-magnet actuator capable of 20-Hz contact-cycle frequency. Their 'mark 2' model (Fig. 1), which works with a 'piezoelectric' actuator, can achieve a frequency of 300 Hz.

But contact-cycle frequency may not be the only way to increase the X-ray yield of Hird and colleagues' apparatus. The triboelectric series and literature suggest that material pairs exist for which contact or frictional electrification leads to charge densities of 10^{13} electrons per square centimetre[8,9]. If fully discharged, such densities would lead to a 1,000-fold increase in X-ray yield to 10^8 photons per stroke or 10^{11} per second at a stroke rate of 1 kHz (for the same contact area of 65 mm^2).

Triboluminescence has been shown[10] to work on microscopic scales, which suggests that, in principle, the device[2] could be scaled down to submillimetre sizes. The challenge will be to manufacture miniature actuators capable of two-dimensional stroke motion for optimal frictional contact. It is possible to imagine a matrix of tiny, individually addressable X-ray sources coupled and synchronized to a fast-readout camera, which would harness the emission from many sources and build up X-ray images in short exposures. If manufacturing techniques for micrometre-sized electromechanical systems can be used, such a triboelectric X-ray source could be realized economically and scaled to large areas (cm^2). Together with their industrial partners, Hird and colleagues have started pursuing this idea. Their work paves the way to a mechanically driven X-ray source for imaging applications in medicine, industry and the life sciences, without the need for a high-voltage power supply. ■

Stefan Kneip *is at the Blackett Laboratory, Imperial College London, London SW7 2BZ, UK.*
e-mail: stefan.kneip@imperial.ac.uk

1. Camara, C. G., Escobar, J. V., Hird, J. R. & Putterman, S. J. *Nature* **455,** 1089–1092 (2008).
2. Hird, J. R., Camara, C. G. & Putterman, S. J. *Appl. Phys. Lett.* **98,** 133501 (2011).
3. Harvey, E. N. *Science* **89,** 460–461 (1939).
4. Walton, A. J. & Reynolds, G. T. *Adv. Phys.* **33,** 595–660 (1984).
5. Walton, A. J. *Adv. Phys.* **26,** 887–948 (1977).
6. Hauksbee, F. *Physico-Mechanical Experiments on Various Subjects* (R. Brugis, 1709).
7. Harper, W. R. *Contact and Frictional Electrification* (Oxford Univ. Press, 1967).
8. Shaw, P. E. *Proc. R. Soc. Lond. A* **94,** 16–33 (1917).
9. Horn, R. G. & Smith, D. T. *Science* **256,** 362–364 (1992).
10. Camara, C. G., Escobar, J. V., Hird, J. R. & Putterman, S. J. *Appl. Phys. B* **99,** 613–617 (2010).

VACCINOLOGY

Persistence pays off

Developing AIDS vaccines has been a frustrating business. A vaccine that triggers immune responses that effectively control early infection by the simian counterpart of HIV in macaques seems promising. SEE LETTER P.523

R. PAUL JOHNSON

HIV is a highly mutable virus that has evolved over millennia to escape host control[1]. It is not surprising, therefore, that researchers have faced numerous challenges in inducing effective responses by the T cells and B cells of the immune system against this virus[2]. Over the past decade, considerable effort has gone into developing AIDS vaccines designed to induce T-cell responses that slow disease progression; such vaccines, however, are unlikely to prevent the explosive burst of viral replication that occurs during primary infection[2]. On page 523 of this issue, Hansen *et al.*[3] describe an alternative approach. They report that in rhesus macaques, the induction of a distinct type of T cell (the effector memory T cell) may limit the early stages of replication of SIV — a virus related to HIV that infects monkeys.

The immune system can recognize pathogens years after initial exposure using a specialized population of T cells called memory T cells. These cells can be divided into two subsets: effector memory T (T_{EM}) cells and central memory T (T_{CM}) cells[4] (Table 1). T_{EM} cells patrol effector sites such as mucosal tissues — the main port of entry for most infectious pathogens, including HIV — and can rapidly kill the infected cells. Efficient maintenance of T_{EM} cells requires continuous antigen stimulation. For example, cytomegalovirus (CMV), a herpesvirus that sets up a persisting infection, is well documented as inducing high-frequency T_{EM}-cell responses[5]. By contrast, T_{CM} cells are primarily found in secondary lymphoid tissues such as lymph nodes and spleen, and are elicited by non-persisting pathogens or vaccines that provide only transient antigenic stimulation[4,6].

How efficiently T_{EM} and T_{CM} cells can

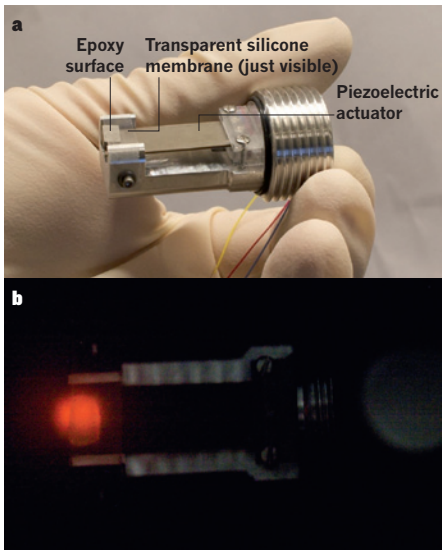


Figure 1 | A triboelectric X-ray machine. **a**, Hird and colleagues' hand-held device, shown here without its vacuum encapsulation, produces X-rays (not shown) when a piezoelectric actuator repeatedly brings an epoxy surface in and out of contact with a silicone membrane. **b**, The image shows the vacuum-enclosed apparatus in operation in a low-pressure neon atmosphere, in which the characteristic orange–red of neon glow discharge can be clearly seen. (Images courtesy of UCLA.)

the ambient gas that surrounds the vacuum-enclosed apparatus should be ionized. This gas ionization is confirmed by the characteristic orange–red of neon glow discharge that is seen when the X-ray source is operated in a low-pressure neon environment (Fig. 1b).

The authors chose silicone because of its strong tendency to charge negatively; a list known as the triboelectric series[8] exists that ranks a material according to its propensity to become charged. To test their device, they added silver to the epoxy surface, and observed the characteristic X-ray K-line emissions (around 22–25 keV) of silver. In doing so, Hird *et al.* prove conclusively that electrons from the discharge process are accelerated to tens of kiloelectronvolts of energy and produce X-rays on impact with the epoxy. The capability to load epoxy with materials of different atomic number gives the apparatus flexibility — it can both be tuned to a desired X-ray line-emission energy and increase the efficiency with which X-rays are generated.

What's more, Hird and colleagues[2] find that, by reducing the ambient pressure, the X-ray emission can persist for more than a second after the epoxy and silicone have been separated. However, to increase the X-ray yield with contact-cycle frequency, it is more desirable to achieve short bursts of X-ray emission. Although this is at the expense of X-ray energy, the emission time can be reduced to less than 10 milliseconds when the device is operated at a higher ambient pressure of 30 millitorr of nitrogen. At such pressure, the authors were able to demonstrate linear scaling of photon number with the frequency of contact cycles. This scaling suggests that the limiting factor to achieving a photon yield of 10^8 per second is finding a linear actuator capable of millimetre displacement and a contact-cycle frequency of 0.1–1 kHz. Their 'mark 1' device was based on a solenoid-magnet actuator capable of 20-Hz contact-cycle frequency. Their 'mark 2' model (Fig. 1), which works with a 'piezoelectric' actuator, can achieve a frequency of 300 Hz.

But contact-cycle frequency may not be the only way to increase the X-ray yield of Hird and colleagues' apparatus. The triboelectric series and literature suggest that material pairs exist for which contact or frictional electrification leads to charge densities of 10^{13} electrons per square centimetre[8,9]. If fully discharged, such densities would lead to a 1,000-fold increase in X-ray yield to 10^8 photons per stroke or 10^{11} per second at a stroke rate of 1 kHz (for the same contact area of 65 mm^2).

Triboluminescence has been shown[10] to work on microscopic scales, which suggests that, in principle, the device[2] could be scaled down to submillimetre sizes. The challenge will be to manufacture miniature actuators capable of two-dimensional stroke motion for optimal frictional contact. It is possible to imagine a matrix of tiny, individually addressable X-ray sources coupled and synchronized to a fast-readout camera, which would harness the emission from many sources and build up X-ray images in short exposures. If manufacturing techniques for micrometre-sized electromechanical systems can be used, such a triboelectric X-ray source could be realized economically and scaled to large areas (cm^2). Together with their industrial partners, Hird and colleagues have started pursuing this idea. Their work paves the way to a mechanically driven X-ray source for imaging applications in medicine, industry and the life sciences, without the need for a high-voltage power supply. ■

Stefan Kneip *is at the Blackett Laboratory, Imperial College London, London SW7 2BZ, UK.*
e-mail: stefan.kneip@imperial.ac.uk

1. Camara, C. G., Escobar, J. V., Hird, J. R. & Putterman, S. J. *Nature* **455,** 1089–1092 (2008).
2. Hird, J. R., Camara, C. G. & Putterman, S. J. *Appl. Phys. Lett.* **98,** 133501 (2011).
3. Harvey, E. N. *Science* **89,** 460–461 (1939).
4. Walton, A. J. & Reynolds, G. T. *Adv. Phys.* **33,** 595–660 (1984).
5. Walton, A. J. *Adv. Phys.* **26,** 887–948 (1977).
6. Hauksbee, F. *Physico-Mechanical Experiments on Various Subjects* (R. Brugis, 1709).
7. Harper, W. R. *Contact and Frictional Electrification* (Oxford Univ. Press, 1967).
8. Shaw, P. E. *Proc. R. Soc. Lond. A* **94,** 16–33 (1917).
9. Horn, R. G. & Smith, D. T. *Science* **256,** 362–364 (1992).
10. Camara, C. G., Escobar, J. V., Hird, J. R. & Putterman, S. J. *Appl. Phys. B* **99,** 613–617 (2010).

VACCINOLOGY

Persistence pays off

Developing AIDS vaccines has been a frustrating business. A vaccine that triggers immune responses that effectively control early infection by the simian counterpart of HIV in macaques seems promising. SEE LETTER P.523

R. PAUL JOHNSON

HIV is a highly mutable virus that has evolved over millennia to escape host control[1]. It is not surprising, therefore, that researchers have faced numerous challenges in inducing effective responses by the T cells and B cells of the immune system against this virus[2]. Over the past decade, considerable effort has gone into developing AIDS vaccines designed to induce T-cell responses that slow disease progression; such vaccines, however, are unlikely to prevent the explosive burst of viral replication that occurs during primary infection[2]. On page 523 of this issue, Hansen *et al.*[3] describe an alternative approach. They report that in rhesus macaques, the induction of a distinct type of T cell (the effector memory T cell) may limit the early stages of replication of SIV — a virus related to HIV that infects monkeys.

The immune system can recognize pathogens years after initial exposure using a specialized population of T cells called memory T cells. These cells can be divided into two subsets: effector memory T (T_{EM}) cells and central memory T (T_{CM}) cells[4] (Table 1). T_{EM} cells patrol effector sites such as mucosal tissues — the main port of entry for most infectious pathogens, including HIV — and can rapidly kill the infected cells. Efficient maintenance of T_{EM} cells requires continuous antigen stimulation. For example, cytomegalovirus (CMV), a herpesvirus that sets up a persisting infection, is well documented as inducing high-frequency T_{EM}-cell responses[5]. By contrast, T_{CM} cells are primarily found in secondary lymphoid tissues such as lymph nodes and spleen, and are elicited by non-persisting pathogens or vaccines that provide only transient antigenic stimulation[4,6].

How efficiently T_{EM} and T_{CM} cells can

TABLE 1 | SUBSETS OF MEMORY T CELLS

	Central memory T (T_{CM}) cells	Effector memory T (T_{EM}) cells
Antigenic stimulation	Transient	Persistent
Triggering vaccines	rAd, DNA, poxviruses	rCMV
Localization	Secondary lymphoid tissues	Mucosal tissues
Markers		
CCR7	++	–
CD127	++	–
CD28	++	–
Perforin	+/–	+++
Proliferative capacity	+++	+/–
Protection against		
Virus acquisition	+/–	+++
Viraemia	++	–

Positive and negative signs denote extent of response/expression. T_{CM} cells and T_{EM} cells vary in different respects: the levels of antigenic stimulation they require, the vaccines that activate them, the molecules they express, the extent of their proliferation and their function. rAd, recombinant adenovirus; rCMV, recombinant cytomegalovirus.

prevent or control an infection varies depending on the pathogen. For HIV and SIV infections, there is no clear consensus on whether T_{EM} or T_{CM} cells are likely to be more effective. Nonetheless, most candidate AIDS vaccines designed to induce T-cell responses that have been tested so far are non-persisting and induce killer ($CD8^+$) T_{CM} cells. These cells can decrease viral load but are relatively ineffective in protecting against initial SIV infection[2].

On the basis of the distinctive characteristics of T_{EM} cells, including their preferential localization to mucosal sites, Hansen and colleagues hypothesized that virus-specific T_{EM} cells might limit SIV replication during a window of vulnerability, in the initial days of infection when relatively few helper ($CD4^+$) T cells are infected[7]. Indeed, this group previously demonstrated[8] that RhCMV/SIV — a recombinant rhesus CMV vaccine expressing multiple SIV proteins — could reduce the risk of progressive infection following repeated rectal SIV administration, most probably by inducing relatively high levels of $CD4^+$ and $CD8^+$ T_{EM} cells.

In their new work[3], Hansen *et al.* extend this approach to a larger cohort of animals. In addition, they examine the effects of a conventional DNA/recombinant adenovirus vaccine that predominantly induces a T_{CM}-cell response. And they give a third group of animals the RhCMV/SIV vaccine first and then boost them with the adenovirus vaccine.

Following repeated rectal challenges of vaccinated and control animals with SIV, 13 of 24 animals that received the RhCMV/SIV vaccine showed a distinctive pattern of transient increase in SIV levels in their blood (viraemia) followed by a rapid decay in the virus levels and then periodic blips of viraemia that became less frequent over the ensuing year. This pattern of transient viraemia(s) stands in stark contrast to the typical course of persistent, high-level viral replication generally observed after SIV infection of macaques.

The animals that could 'control' SIV infection also developed new or increased T-cell responses to the viral protein Vif, which confirmed that a limited 'take' of SIV infection had occurred in these animals. The authors, however, could not detect SIV at post-mortem examination in a subset of animals with transient SIV infection, despite intensive efforts. Together, these findings suggest that induction of virus-specific T_{EM} cells may control productive SIV replication in a presently undefined reservoir of infected cells before the development of progressive, systemic infection.

How can one tell that T_{EM} cells were responsible for controlling the establishment of progressive SIV infection in these animals? For reasons that are not yet clear, RhCMV vectors do not efficiently induce antibody responses to SIV proteins, including viral envelope proteins. So B cells are unlikely to play a part. Moreover, the SIV-specific $CD4^+$ and $CD8^+$ T cells in RhCMV/SIV-vaccinated animals had a characteristic T_{EM} phenotype and were enriched in effector sites such as the gut mucosa.

Hansen *et al.* report that, compared with animals that had progressive SIV infection, those that did not develop such infection had higher peak frequencies of SIV-specific $CD8^+$ T cells after immunization. This suggests that the extent to which mucosal sites are 'seeded' with T_{EM} cells during immunization may be a crucial factor in subsequent control of infection. It is essential to determine whether this unusual pattern of viral control can be observed in an independent cohort of animals vaccinated with RhCMV/SIV and to determine which T-cell populations mediate the protective effect. Whether protection occurs when vaccinated macaques are exposed to SIV vaginally or intravenously also remains to be seen.

It is noteworthy that just over half of the animals Hansen *et al.* vaccinated with RhCMV/SIV displayed a pattern of controlled infection, and once infected, the RhCMV/SIV-vaccinated animals showed no better control of viraemia than unvaccinated controls. So if T_{EM} cells pan out to be an important component of an AIDS vaccine strategy, optimally they would be combined with vaccines that induce both neutralizing antibodies (assuming the considerable challenges in inducing these antibodies can be addressed[2]) and T_{CM} cells capable of controlling viraemia in the event of systemic infection. No matter how desirable, the combined induction of T_{CM} and T_{EM} cells is not trivial: continual antigenic stimulation may deplete the T_{CM}-cell pool over time.

Can induction of T_{EM} cells by CMV-based vectors serve as a viable HIV vaccine for humans? In healthy hosts, CMV can cause a disease similar to mononucleosis. And in immunocompromised patients as well as infants with congenital infection it can lead to severe disease. Clinical use of CMV vectors is therefore likely to encounter significant scrutiny.

A clinically acceptable CMV vaccine ought to be non-pathogenic in seronegative hosts. It must also present a minimal risk of mother-to-child transmission and of genital shedding (to minimize inadvertent transmission to others), while maintaining immunogenicity and persistence. The expected trade-offs between attenuation of pathogenicity and immunogenicity have already been documented[9] in limited clinical trials of live recombinant CMV vaccines. But work is under way to develop recombinant viruses that are attenuated in pathogenicity and retain the immunogenicity of the unmodified CMV vaccines. For the future of CMV vectors, as well as for the overall AIDS vaccine enterprise, ultimately, persistence may well pay off. ■

R. Paul Johnson *is in the Division of Immunology, New England Primate Research Center, Harvard Medical School, Southborough, Massachusetts 01772, USA. He is also at the Ragon Institute of Massachusetts General Hospital, MIT and Harvard, and the Infectious Disease Unit, Massachusetts General Hospital, Boston.*
e-mail: paul_johnson@hms.harvard.edu

1. Johnson, W. E. & Desrosiers, R. C. *Annu. Rev. Med.* **53,** 499–518 (2002).
2. Barouch, D. H. *Nature* **455,** 613–619 (2008).
3. Hansen, S. G. *et al. Nature* **473,** 523–527 (2011).
4. Sallusto, F., Mackay, C. R. & Lanzavecchia, A. *Annu. Rev. Immunol.* **18,** 593–620 (2000).
5. Moss, P. & Khan, N. *Hum. Immunol.* **65,** 456–464 (2004).
6. Robinson, H. L. & Amara, R. R. *Nature Med.* **11,** S25–S32 (2005).
7. Haase, A. T. *Nature* **464,** 217–223 (2010).
8. Hansen, S. G. *et al. Nature Med.* **15,** 293–299 (2009).
9. Heineman, T. C. *et al. J. Infect. Dis.* **193,** 1350–1360 (2006).

MEDICINAL CHEMISTRY

New lead for pain treatment

The synthesis of conolidine, a scarce, naturally occurring compound, has enabled the first studies of its pharmacological properties to be carried out. Excitingly, conolidine is a painkiller that seems to have an unusual mechanism of action.

SARAH E. REISMAN

Some of the most powerful painkillers, such as morphine, hydrocodone and oxycodone, belong to the opioid family of analgesics. Unfortunately, prolonged exposure to these analgesics can cause several adverse side effects, including physical and psychological addiction. As a result, there is a continued effort to identify painkillers that have different biological mechanisms from opioids and that elicit fewer side effects. With this in mind, a team of chemists, led by Glenn Micalizio, has joined forces with a group of neuroscientists, headed by Laura Bohn, to synthesize and study the analgesic properties of a rare, naturally occurring compound called conolidine. Their promising findings, reported in *Nature Chemistry* (Tarselli *et al.*[1]), might pave the way for the development of new non-opioid analgesics.

The natural product conolidine was originally isolated[2] in extremely small quantities — just 0.00014% yield — from the stem bark of the flowering tropical plant *Tabernaemontana divaricata*. The low natural abundance of the compound has hindered the study of its potential therapeutic properties. By completing the first total chemical synthesis of conolidine, Micalizio and co-workers provided Bohn's team with enough synthetic material to carry out the first *in vivo* studies of its analgesic properties.

A key challenge in the synthesis of conolidine is the construction of its bicyclic ring system (Fig. 1a), which consists of an eight-membered ring bridged by two carbon atoms and contains a nitrogen atom at one of the bridgehead positions. In devising a strategy to prepare this system, the authors turned to nature for inspiration.

Conolidine is a 'C5-nor stemmadenine' natural product, which means that it contains the same carbon skeleton as the more abundant natural product stemmadenine, except that it lacks one of the carbon atoms (known as C5; see Fig. 1b). It has been proposed[3] that, in the biosynthesis of C5-nor stemmadenine compounds, the C5 atom of a stemmadenine framework is excised in a process that begins with the oxidation of the bridgehead nitrogen (Fig. 1c). This proposal has been validated chemically — stemmadenine can be converted to its C5-nor analogue vallesamine in this manner, albeit in low (25%) yield[4].

Rather than pursuing a strictly biomimetic sequence in which a compound containing the stemmadenine framework was converted to conolidine, Micalizio and the chemistry team[1] adopted a more subtle approach. They developed a short synthesis of a compound that has all but one of the carbons found in conolidine (Fig. 1d). They then incorporated the final carbon by reacting the compound with formaldehyde (CH_2O), generating a precursor similar to that produced in the biomimetic synthesis of vallesamine. This precursor then underwent an intramolecular cyclization reaction to forge the critical eight-membered ring of the bicyclic system.

This approach enabled the team to prepare conolidine in a more straightforward manner than would have been possible by first making the stemmadenine framework and then fragmenting it. In addition, using this approach, they were able to independently prepare not only the naturally occurring isomer of conolidine, but also its unnatural mirror-image isomer (known as (–)-conolidine), in only ten synthetic steps from a commercially available starting material.

With access to synthetic conolidine, Bohn and her group[1] went on to evaluate the compound's analgesic properties in mice. In experiments designed to evaluate conolidine's effects on both acute and persistent pain, they found that it was indeed a painkiller of similar potency to morphine. However, the authors' pharmacological studies revealed that conolidine does not bind to opioid receptors (the biological targets responsible for the analgesic effects of morphine and other opioid drugs). Furthermore, the authors found that conolidine does not seem to adversely affect the locomotor activity of mice, as opioid analgesics do, suggesting that conolidine might have fewer side effects than opioids.

Intriguingly, Bohn and colleagues also observed that (–)-conolidine has comparable *in vivo* activities to (+)-conolidine, the naturally occurring isomer. This is unusual, because the two mirror-image isomers (enantiomers) of a compound commonly elicit

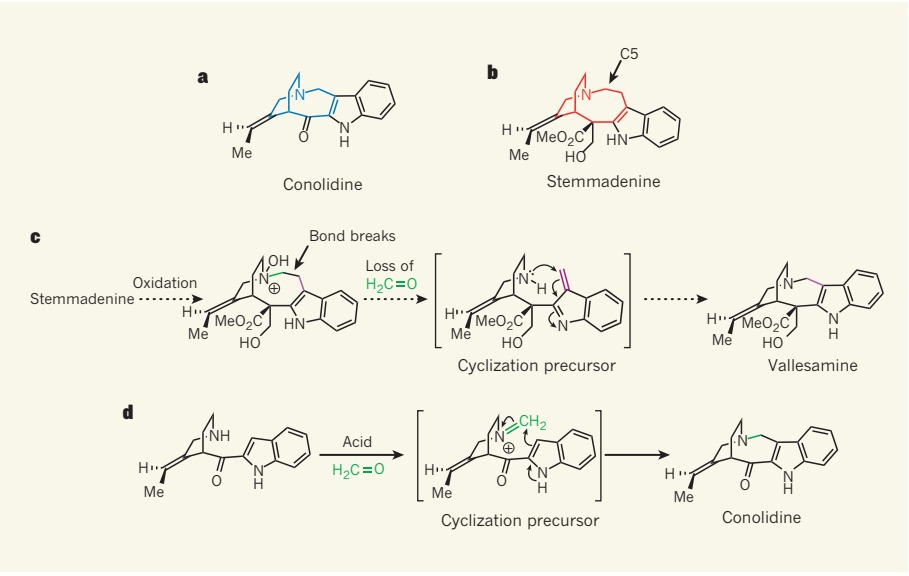


Figure 1 | Inspiration for the synthesis of conolidine. **a**, Conolidine is a scarce, plant-derived compound. Its core structure is shown in blue. Me is a methyl group. **b**, Stemmadenine is a related natural product. Its core structure (red) has one more carbon atom than conolidine; the carbon atom is known as C5. **c**, The biosyntheses of compounds containing the core structure of conolidine are thought to involve a process in which the C5 carbon is excised from the stemmadenine core. In this example, the oxidation of a nitrogen atom in stemmadenine triggers the loss of C5 as formaldehyde (CH_2O) and generates a cyclization precursor, which undergoes an intramolecular reaction to yield vallesamine as a product. Curly arrows show the electron movement in the intramolecular reaction. **d**, In their synthesis of conolidine, Tarselli *et al.*[1] prepared a compound that, on treatment with an acid and formaldehyde, formed a cyclization precursor similar to that shown in **c**. The precursor underwent an intramolecular reaction that formed the desired product.

different biological responses. For example, the enantiomer of morphine is a poor painkiller[5]. The fact that both enantiomers of conolidine are analgesic may indicate something about its biological target. Taken together, the biological findings suggest that conolidine may have a previously undiscovered pharmacological mechanism for inducing analgesia.

Just what the mechanism of action is has not been determined, and this is clearly a priority for future research. Micalizio, Bohn and colleagues' preliminary studies[1] nevertheless indicate that conolidine is a promising candidate for further study as a non-opioid analgesic. Moreover, the authors' concise, modular and high-yielding chemical synthesis should provide ample quantities of conolidine for the further study and development of this painkiller — quantities that would be extremely difficult to extract from the natural source of the compound. ■

Sarah E. Reisman *is in the Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, USA.*
e-mail: reisman@caltech.edu

1. Tarselli, M. A. *et al. Nature Chem.* **3,** 449–453 (2011).
2. Kam, T.-S., Pang, H.-S., Choo, Y.-M. & Komiyama, K. *Chem. Biodiver.* **1,** 646–656 (2004).
3. Potier, P. & Janot, M. M. *C.R. Acad. Sci.* **276C,** 1727 (1973).
4. Scott, A. I., Yeh, C.-L. & Greenslade, D. *J. Chem. Soc. Chem. Commun.* 947–948 (1978).
5. Rice, K. C. in *The Chemistry and Biology of Isoquinoline Alkaloids* (eds Phillipson, J. D., Roberts, M. F. & Zenk, M. H.) 191–203 (Springer, 1985).

PRECISION MEASUREMENT

A search for electrons that do the twist

One might think that physicists know everything about the electron. But the latest measurement of its shape could alter expectations for results at high-energy particle accelerators. SEE LETTER P.493

AARON E. LEANHARDT

If I were to tell you about an elementary particle that has mass and charge, but neither size nor structure, yet still has a well-defined orientation and can point in a specific direction in space, you would probably think I am describing something from a science-fiction novel. In fact, I am telling you about the electron. On page 493 of this issue, Hudson *et al.*[1] describe an experiment aimed at refining our understanding of this fundamental particle and, more broadly, the basic laws of nature.

Described colloquially, their experiment searches for evidence of an aspheric distortion to the shape of the electron, or, more technically, to the shape of its interactions with electric fields. Hudson *et al.*[1] observe no such distortion. However, a detailed understanding of their apparatus allows them to report their null result as a new limit on the magnitude of the electric dipole moment of the electron. This work has important ramifications for the types of particles that can be discovered at high-energy accelerators, and may eventually help to explain the composition of the observable Universe.

It is well established that the electron has a magnetic dipole moment, which means that it behaves like a tiny bar magnet with north and south poles. For example, a magnetic field can rotate the orientation of an electron, just as it can move the needle of a compass. Hudson *et al.*[1] are searching for the as-yet undiscovered electric analogue, the electric dipole moment of the electron. An electric dipole moment can be depicted as a battery with positive and negative terminals, and its orientation can be rotated by electric fields. Therefore, the experimental effort of Hudson and colleagues can be viewed as an attempt to answer the question: does an electric field twist the orientation of an electron?

It should be expected that stronger electric fields and longer measurement times would enhance the probability of observing the electron 'doing the twist'. Herein lies the difficulty. A free electron will accelerate under the influence of an electric field and crash into the walls of the apparatus. This effect is extremely useful for generating X-rays in medical devices and security scanners, but in the present experiment it has only the detrimental effect of limiting the measurement time. This obstacle can be overcome by binding several electrons to a heavy nucleus to form a neutral atom comprising a central core and some outer valence electrons. An electric field will not accelerate this neutral atom, but will polarize it — that is, it will separate opposite charges within the atom. Furthermore, the effective electric field 'seen' by the valence electrons in a suitably chosen and properly polarized neutral atom can be quite large[2]. The previous best attempt to detect the electric dipole moment of the electron was made by probing the valence electrons in a beam of neutral thallium atoms[3].

Even before the thallium-based experiment[3] was completed, techniques to improve on it were being devised. Molecules are typically easier to polarize than atoms, which translates into the molecular valence electrons experiencing even larger effective electric fields[4,5]. This benefit was crucial in enabling Hudson *et al.*[1], who worked with ytterbium monofluoride (YbF), to surpass the measurement sensitivity achieved in the thallium-based experiment[3] — albeit, at present, by a modest factor of 1.5. Specifically, the authors limit the magnitude of the electron's electric dipole moment to less than 10.5×10^{-28} e centimetres, where e is the charge of the electron. In electrostatic units, this value is more than 16 orders of magnitude weaker than the known magnetic dipole moment of the electron. Hudson and colleagues have pioneered the use of cold polar molecules to push the search for an electric dipole moment of the electron to new levels, and their work serves as a gateway to multiple next-generation molecule-based experiments. These experiments[6–10], as well as a continued effort by Hudson *et al.*[1], are aiming to improve on the above-mentioned limit by a factor of 10–100.

How can studying a sizeless and structureless particle be so interesting? The interest arises from its interaction with another seemingly featureless entity — empty space, casually called the vacuum. In reality, empty space is not always so empty. The vacuum comprises a sea of particles that are hopping into and out of existence like waves crashing onto a shore and then receding back to the ocean. These whimsical particles do not stick around for long enough to be observed directly. However, they make their presence felt through their interaction with commonplace matter, such as the electrons studied by Hudson and colleagues[1].

Physicists contend that it is these particles that give the electron its electric dipole moment, almost as if they are the band playing just the right music required for the electrons to do the twist. Without these particles, no electric field would be strong enough and no measurement time long enough for us to see the electrons dance. Furthermore, they are a subset of the new particles that physicists working at high-energy accelerators are hoping to create and observe directly. Hence, searches for the electric dipole moment of the electron provide crucial information about phenomena that naturally occur at energies 10^{30} times greater than those directly measured in the precision tabletop work of Hudson *et al.*[1].

In 1950, common theoretical arguments asserted that fundamental particles could not

different biological responses. For example, the enantiomer of morphine is a poor painkiller[5]. The fact that both enantiomers of conolidine are analgesic may indicate something about its biological target. Taken together, the biological findings suggest that conolidine may have a previously undiscovered pharmacological mechanism for inducing analgesia.

Just what the mechanism of action is has not been determined, and this is clearly a priority for future research. Micalizio, Bohn and colleagues' preliminary studies[1] nevertheless indicate that conolidine is a promising candidate for further study as a non-opioid analgesic. Moreover, the authors' concise, modular and high-yielding chemical synthesis should provide ample quantities of conolidine for the further study and development of this painkiller — quantities that would be extremely difficult to extract from the natural source of the compound. ■

Sarah E. Reisman *is in the Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, USA.*
e-mail: reisman@caltech.edu

1. Tarselli, M. A. *et al. Nature Chem.* **3,** 449–453 (2011).
2. Kam, T.-S., Pang, H.-S., Choo, Y.-M. & Komiyama, K. *Chem. Biodiver.* **1,** 646–656 (2004).
3. Potier, P. & Janot, M. M. *C.R. Acad. Sci.* **276C,** 1727 (1973).
4. Scott, A. I., Yeh, C.-L. & Greenslade, D. *J. Chem. Soc. Chem. Commun.* 947–948 (1978).
5. Rice, K. C. in *The Chemistry and Biology of Isoquinoline Alkaloids* (eds Phillipson, J. D., Roberts, M. F. & Zenk, M. H.) 191–203 (Springer, 1985).

PRECISION MEASUREMENT

A search for electrons that do the twist

One might think that physicists know everything about the electron. But the latest measurement of its shape could alter expectations for results at high-energy particle accelerators. SEE LETTER P.493

AARON E. LEANHARDT

If I were to tell you about an elementary particle that has mass and charge, but neither size nor structure, yet still has a well-defined orientation and can point in a specific direction in space, you would probably think I am describing something from a science-fiction novel. In fact, I am telling you about the electron. On page 493 of this issue, Hudson *et al.*[1] describe an experiment aimed at refining our understanding of this fundamental particle and, more broadly, the basic laws of nature.

Described colloquially, their experiment searches for evidence of an aspheric distortion to the shape of the electron, or, more technically, to the shape of its interactions with electric fields. Hudson *et al.*[1] observe no such distortion. However, a detailed understanding of their apparatus allows them to report their null result as a new limit on the magnitude of the electric dipole moment of the electron. This work has important ramifications for the types of particles that can be discovered at high-energy accelerators, and may eventually help to explain the composition of the observable Universe.

It is well established that the electron has a magnetic dipole moment, which means that it behaves like a tiny bar magnet with north and south poles. For example, a magnetic field can rotate the orientation of an electron, just as it can move the needle of a compass. Hudson *et al.*[1] are searching for the as-yet undiscovered electric analogue, the electric dipole moment of the electron. An electric dipole moment can be depicted as a battery with positive and negative terminals, and its orientation can be rotated by electric fields. Therefore, the experimental effort of Hudson and colleagues can be viewed as an attempt to answer the question: does an electric field twist the orientation of an electron?

It should be expected that stronger electric fields and longer measurement times would enhance the probability of observing the electron 'doing the twist'. Herein lies the difficulty. A free electron will accelerate under the influence of an electric field and crash into the walls of the apparatus. This effect is extremely useful for generating X-rays in medical devices and security scanners, but in the present experiment it has only the detrimental effect of limiting the measurement time. This obstacle can be overcome by binding several electrons to a heavy nucleus to form a neutral atom comprising a central core and some outer valence electrons. An electric field will not accelerate this neutral atom, but will polarize it — that is, it will separate opposite charges within the atom. Furthermore, the effective electric field 'seen' by the valence electrons in a suitably chosen and properly polarized neutral atom can be quite large[2]. The previous best attempt to detect the electric dipole moment of the electron was made by probing the valence electrons in a beam of neutral thallium atoms[3].

Even before the thallium-based experiment[3] was completed, techniques to improve on it were being devised. Molecules are typically easier to polarize than atoms, which translates into the molecular valence electrons experiencing even larger effective electric fields[4,5]. This benefit was crucial in enabling Hudson *et al.*[1], who worked with ytterbium monofluoride (YbF), to surpass the measurement sensitivity achieved in the thallium-based experiment[3] — albeit, at present, by a modest factor of 1.5. Specifically, the authors limit the magnitude of the electron's electric dipole moment to less than 10.5×10^{-28} e centimetres, where e is the charge of the electron. In electrostatic units, this value is more than 16 orders of magnitude weaker than the known magnetic dipole moment of the electron. Hudson and colleagues have pioneered the use of cold polar molecules to push the search for an electric dipole moment of the electron to new levels, and their work serves as a gateway to multiple next-generation molecule-based experiments. These experiments[6–10], as well as a continued effort by Hudson *et al.*[1], are aiming to improve on the above-mentioned limit by a factor of 10–100.

How can studying a sizeless and structureless particle be so interesting? The interest arises from its interaction with another seemingly featureless entity — empty space, casually called the vacuum. In reality, empty space is not always so empty. The vacuum comprises a sea of particles that are hopping into and out of existence like waves crashing onto a shore and then receding back to the ocean. These whimsical particles do not stick around for long enough to be observed directly. However, they make their presence felt through their interaction with commonplace matter, such as the electrons studied by Hudson and colleagues[1].

Physicists contend that it is these particles that give the electron its electric dipole moment, almost as if they are the band playing just the right music required for the electrons to do the twist. Without these particles, no electric field would be strong enough and no measurement time long enough for us to see the electrons dance. Furthermore, they are a subset of the new particles that physicists working at high-energy accelerators are hoping to create and observe directly. Hence, searches for the electric dipole moment of the electron provide crucial information about phenomena that naturally occur at energies 10^{30} times greater than those directly measured in the precision tabletop work of Hudson *et al.*[1].

In 1950, common theoretical arguments asserted that fundamental particles could not

have electric dipole moments. But Purcell and Ramsey[11] realized at the time that such arguments were based on untested assumptions, and declared: "The question of the possible existence of an electric dipole moment of a nucleus or of an elementary particle in view of the above becomes a purely experimental matter."

Today, typical theories predict electric dipole moments for many fundamental particles, including the electron, but the predictions span a wide range of values. Therefore, despite the complete reversal of opinion on the theoretical front, the essence of Purcell and Ramsey's claim endures. Establishing the existence of an electric dipole moment of a fundamental particle is an exclusively experimental endeavour. Hudson *et al.*[1] are the latest to attempt such a feat. Experiments of this genre reach far beyond the realm of atomic, molecular and optical physics: they can be viewed as low-energy windows on the high-energy soul of the cosmos. ■

Aaron E. Leanhardt *is in the Department of Physics, University of Michigan, Ann Arbor, Michigan 48109-1040, USA.*
e-mail: aehardt@umich.edu

1. Hudson, J. J. *et al. Nature* **473,** 493–496 (2011).
2. Sandars, P. G. H. *Phys. Lett.* **14,** 194–196 (1965).
3. Regan, B. C., Commins, E. D., Schmidt, C. J. & DeMille, D. *Phys. Rev. Lett.* **88,** 071805 (2002).
4. Sandars, P. G. H. *Phys. Rev. Lett.* **19,** 1396–1398 (1967).
5. Sushkov, O. P. & Flambaum, V. V. *Sov. Phys. JETP* **48,** 608–611 (1978).
6. Alphei, L. D. *et al. Phys. Rev. A* **83,** 040501 (2011).
7. Bickman, S., Hamilton, P., Jiang, Y. & DeMille, D. *Phys. Rev. A* **80,** 023418 (2009).
8. Leanhardt, A. E. *et al.* Preprint at http://arxiv.org/abs/1008.2997 (2010).
9. Vutha, A. C. *et al. J. Phys. B* **43,** 074007 (2010).
10. Lee, J. *et al. J. Mod. Opt.* **56,** 2005–2012 (2009).
11. Purcell, E. M. & Ramsey, N. F. *Phys. Rev.* **78,** 807 (1950).

PLANETARY SCIENCE

Building a planet in record time

It seems that Mars had grown to near its present size by 2 million to 4 million years after the Solar System began to form. Such rapid growth explains why the planet is much smaller than Earth and Venus. SEE LETTER P.489

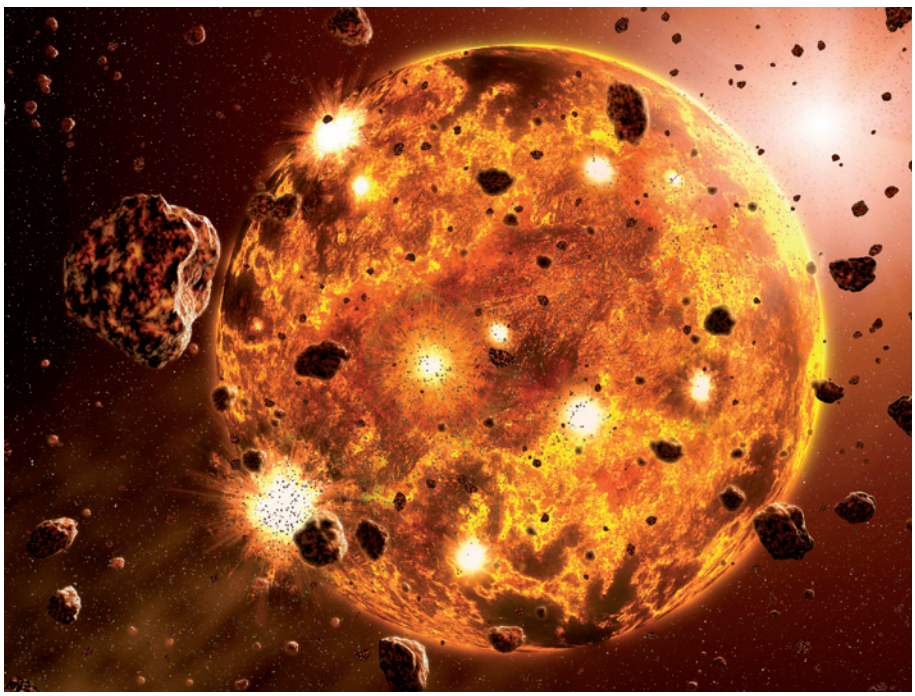
ALAN BRANDON

How long did the rocky planets Mercury, Venus, Earth and Mars take to form? Answering this question will tell us why our planets look the way they do today. Previous estimates[1,2] place the formation of Mars at up to 15 million years from the time the Solar System began to form. On page 489 of this issue, Dauphas and Pourmand[3] derive even tighter constraints on the planet's formation age by determining Mars's abundance ratio of hafnium to tungsten (Hf/W) and then re-evaluating the age obtained using a chronometer based on the decay of ^{182}Hf to ^{182}W.

The amount of ^{182}W in meteorites from Mars can be used to place constraints on its age of formation. The isotope ^{182}Hf decays to ^{182}W with a half-life of 9 million years, and can date events that occurred in the first 60 million years or so of Solar System history, before most ^{182}Hf decayed away. During their early history, rocky planets differentiate into iron-rich metal cores and silicate-rich mantles. Tungsten is siderophile (it likes to bond with iron) and so partitions into the iron-rich cores. Hafnium remains in silicate and oxide minerals (it is lithophile) in the newly formed mantles. Hence, the age of core formation of a planet is recorded in the tungsten isotopic compositions of planetary materials. Core formation is thought to occur at or near the time that planets reach their final mass.

The tungsten isotope compositions of Martian meteorites have been accurately determined. But calculating the age of Mars's core formation also depends on knowing its bulk silicate Hf/W ratio. These meteorites are igneous rocks that were produced by the melting of rock deep within Mars, and that subsequently migrated and cooled near or at its surface. This migration probably resulted in fractionation of Hf and W in the magmas relative to their sources. To better determine the Hf/W ratio of bulk silicate Mars, Dauphas and Pourmand[3] used the fact that the ratio of thorium to tungsten (Th/W) in Martian meteorites is constant, and recognized that the Th/Hf ratio of Mars should not differ from the average bulk Solar System value because of the similar chemical behaviours of Th and Hf in Mars during igneous processing.

Armed with this information, the authors[3] accurately determined the Th/Hf ratio of stony meteorites (chondrites), which represent the average bulk Solar System ratio, and used this as a proxy for the Th/Hf ratio of Mars, from which they calculated its bulk silicate Hf/W ratio. By combining their calculated bulk silicate Mars Hf/W ratio with the W isotopic compositions of Martian meteorites, the authors were able to determine an age of core formation for the planet — a maximum of around 2 million to 4 million years after the Solar System began to form. This rapid formation time explains why Mars is



TAKE 27/SPL

Figure 1 | Planetary accretion. This illustration shows small rocky bodies accreting to a larger body, a protoplanet. Such accretion is thought to be the way in which protoplanets grow to become planets.

averaging over heterogeneous cell populations.

The simplicity of the technique, called SiMPull, is striking, raising the question of how the required specificity and impressive signal-to-noise ratio presented by Jain *et al.* is achieved. The adsorption of nonspecific proteins is minimized by using methoxy polyethylene glycol (mPEG) monolayers on the coverslips. Biotinylated PEG molecules, together with neutravidin, act as anchors for biotinylated antibodies directed against the bait protein (Fig. 1). The authors first validated the system by demonstrating efficient and specific immobilization for a polyhistidine-tagged variant of the yellow fluorescent protein (YFP). The signal-to-noise ratio was maintained at ten or more by adjusting lysate dilution factors.

The sample preparation conditions are also mild. Sensitive protein assemblies, such as intact membrane protein complexes (the β_2-adrenergic receptor), or even membrane patches, were successfully pulled down with similar efficiency and data quality. In addition, the authors show that potential problems arising from the expression of modified protein (for example, altered properties and increased or decreased expression levels compared with the wild-type protein) can be overcome by immunofluorescence detection of only endogenous complexes.

Does this work[1] present a new gold standard for analysing protein–protein interactions? To answer this question, the details of the method have to be considered. The bait protein is captured using a specific antibody or an affinity tag. Once immobilized, the proteins are detected using appropriate antibodies or, alternatively, the signal of a fused fluorescent reporter protein or fluorescent antibody can be recorded. Therefore the method is limited to well-known targets and the screening of new interaction partners is not feasible. Whether it can compete, for example, with label-free techniques such as mass spectrometry that enable the identification of protein–protein interactions with fewer constraints[4,5], is arguable. In this respect SiMPull can be viewed as an extension of western-blot analysis.

To fully appreciate the potential of the work by Jain *et al.*[1], however, the wealth of possible applications beyond the mere detection of a protein–protein interaction has to be considered. Once complexes are immobilized on the imaging surface, the trump cards offered by single-molecule fluorescence microscopy can be played, circumventing the static and dynamic averaging of common biomolecular assays. In an impressive series of examples, the authors demonstrate the variety of information that can be gained using SiMPull.

Proteins can be counted one by one, and protein expression levels can be quantified by comparison with a reference such as a recombinant protein. The stoichiometries of protein complexes can be determined by counting successive bleaching steps of single molecules, as is shown for YFP and tandem dimeric YFP, and other monomeric or dimeric proteins[6]. And multicolour imaging can be used to determine stoichiometries of heterogeneous protein complexes[7–9]. As an example, the authors show that the regulatory and catalytic subunits of the inactive tetrameric protein kinase A (PKA) are pulled down together, and that both domains are immobilized as a complex.

Moreover, immobilized complexes can be challenged in functional assays. Adding the activator cyclic AMP, which induces dissociation of the PKA complex, resulted in greatly reduced numbers of co-localized spots, confirming that constructs retain their properties. Increasing intracellular cAMP levels by external stimuli before performing SiMPull also yielded a reduced number of co-localized catalytic and regulatory domains, showing that subpopulations and changing protein interactions in cells can be revealed.

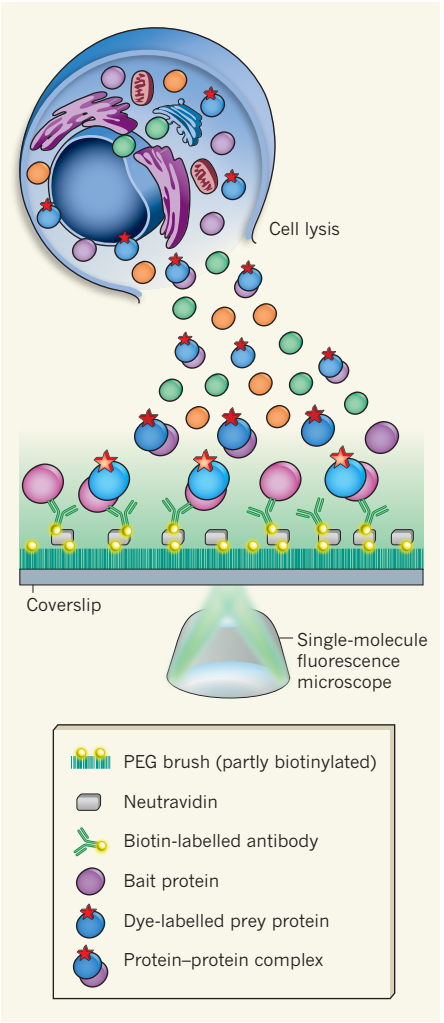


Figure 1 | The workflow of single-molecule pull-down (SiMPull)[1]. The cell lysate is applied directly to the imaging surface for single-molecule fluorescence microscopy. Protein complexes of interest are captured using specific antibodies on the surface. Prey proteins associated with the bait protein can be detected using, for example, a fluorescent dye fused to the prey. PEG, polyethylene glycol.

Above all, Jain *et al.*[1] demonstrate a sophisticated single-molecule FRET (fluorescence resonance energy transfer) experiment using immobilized PcrA helicase as a model protein. Addition of a partial duplex DNA with a 5′ overhang and ATP to the PcrA allowed the real-time observation of the helicase activity through FRET changes in the doubly labelled DNA. The direct single-molecule sample preparation by SiMPull opens up a route to study proteins in their natural complexes that are difficult to reproduce in *in vitro* experiments using recombinant proteins.

SiMPull, then, offers a great deal. It combines the principles of conventional pull-down assays with single-molecule microscopy and enables the direct visualization of cellular protein complexes. Known interactions are revealed in a robust and convenient manner, under mild preparation conditions, thereby circumventing the problems of imaging methods in living cells. SiMPull allows the determination of stoichiometries even for sensitive and short-lived protein complexes, and subpopulations arising from physiological permutations of protein–protein interactions can be revealed.

In the long term, when combined with automated workflows and microfluidics, SiMPull will possibly allow the high-throughput study of variable complex formation as a function of external stimuli such as cell stress. Clever experimental design and optimal use of the information contained in the single-molecule experiment — for example involving FRET and subnanometre localization[10] — might even allow protein pairs that physically interact, and those that happen to be in the same complex, to be distinguished. In the meantime, the numerous applications presented by Jain *et al.* will inspire other researchers, and single-molecule detection might be the key to take other important techniques to the next level. ■

Philip Tinnefeld *is at the Institute for Physical and Theoretical Chemistry, Braunschweig University of Technology, 38106 Braunschweig, Germany.*
e-mail: p.tinnefeld@tu-braunschweig.de

1. Jain, A. *et al. Nature* **473,** 484–488 (2011).
2. Puig, O. *et al. Methods* **24,** 218–229 (2001).
3. Barrios-Rodiles, M. *et al. Science* **307,** 1621–1625 (2005).
4. Vermeulen, M., Hubner, N. C. & Mann, M. *Curr. Opin. Biotechnol.* **19,** 331–337 (2008).
5. Gingras, A. C., Gstaiger, M., Raught, B. & Aebersold, R. *Nature Rev. Mol. Cell Biol.* **8,** 645–654 (2007).
6. Ulbrich, M. H. & Isacoff, E. Y. *Nature Methods* **4,** 319–321 (2007).
7. Kapanidis, A. N. *et al. Proc. Natl Acad. Sci. USA* **101,** 8936–8941 (2004).
8. Lee, J. *et al. Angew. Chem. Int. Edn* **122,** 10118–10121 (2010).
9. Stein, I. H., Steinhauer, C. & Tinnefeld, P. *J. Am. Chem. Soc.* **133,** 4193–4195 (2011).
10. Pertsinidis, A., Zhang, Y. & Chu, S. *Nature* **466,** 647–651 (2010).

have electric dipole moments. But Purcell and Ramsey[11] realized at the time that such arguments were based on untested assumptions, and declared: "The question of the possible existence of an electric dipole moment of a nucleus or of an elementary particle in view of the above becomes a purely experimental matter."

Today, typical theories predict electric dipole moments for many fundamental particles, including the electron, but the predictions span a wide range of values. Therefore, despite the complete reversal of opinion on the theoretical front, the essence of Purcell and Ramsey's claim endures. Establishing the existence of an electric dipole moment of a fundamental particle is an exclusively experimental endeavour. Hudson *et al.*[1] are the latest to attempt such a feat. Experiments of this genre reach far beyond the realm of atomic, molecular and optical physics: they can be viewed as low-energy windows on the high-energy soul of the cosmos. ■

Aaron E. Leanhardt *is in the Department of Physics, University of Michigan, Ann Arbor, Michigan 48109-1040, USA.*
e-mail: aehardt@umich.edu

1. Hudson, J. J. *et al. Nature* **473,** 493–496 (2011).
2. Sandars, P. G. H. *Phys. Lett.* **14,** 194–196 (1965).
3. Regan, B. C., Commins, E. D., Schmidt, C. J. & DeMille, D. *Phys. Rev. Lett.* **88,** 071805 (2002).
4. Sandars, P. G. H. *Phys. Rev. Lett.* **19,** 1396–1398 (1967).
5. Sushkov, O. P. & Flambaum, V. V. *Sov. Phys. JETP* **48,** 608–611 (1978).
6. Alphei, L. D. *et al. Phys. Rev. A* **83,** 040501 (2011).
7. Bickman, S., Hamilton, P., Jiang, Y. & DeMille, D. *Phys. Rev. A* **80,** 023418 (2009).
8. Leanhardt, A. E. *et al.* Preprint at http://arxiv.org/abs/1008.2997 (2010).
9. Vutha, A. C. *et al. J. Phys. B* **43,** 074007 (2010).
10. Lee, J. *et al. J. Mod. Opt.* **56,** 2005–2012 (2009).
11. Purcell, E. M. & Ramsey, N. F. *Phys. Rev.* **78,** 807 (1950).

PLANETARY SCIENCE

Building a planet in record time

It seems that Mars had grown to near its present size by 2 million to 4 million years after the Solar System began to form. Such rapid growth explains why the planet is much smaller than Earth and Venus. SEE LETTER P.489

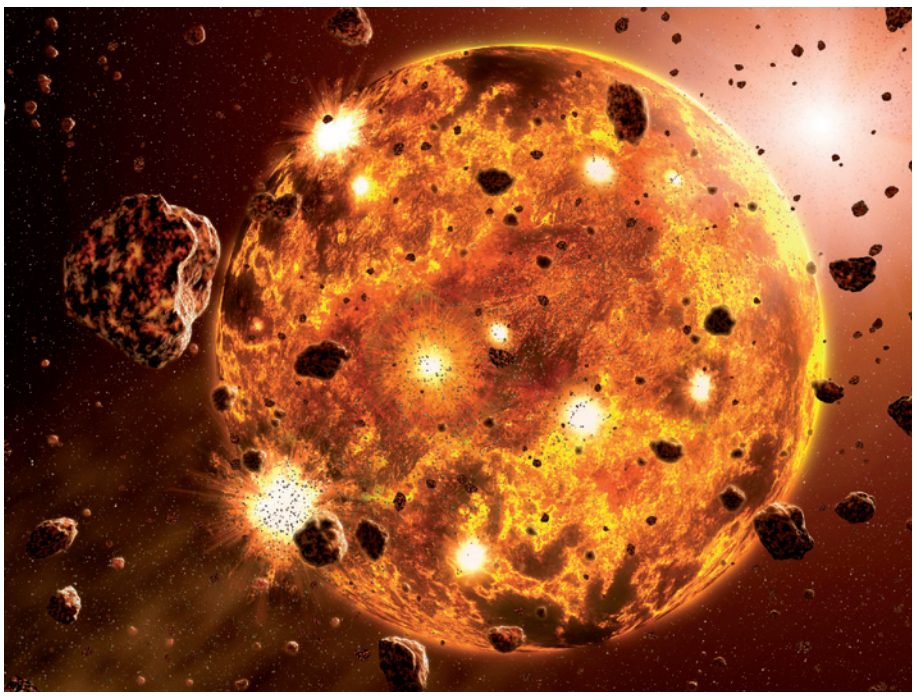
ALAN BRANDON

How long did the rocky planets Mercury, Venus, Earth and Mars take to form? Answering this question will tell us why our planets look the way they do today. Previous estimates[1,2] place the formation of Mars at up to 15 million years from the time the Solar System began to form. On page 489 of this issue, Dauphas and Pourmand[3] derive even tighter constraints on the planet's formation age by determining Mars's abundance ratio of hafnium to tungsten (Hf/W) and then re-evaluating the age obtained using a chronometer based on the decay of ^{182}Hf to ^{182}W.

The amount of ^{182}W in meteorites from Mars can be used to place constraints on its age of formation. The isotope ^{182}Hf decays to ^{182}W with a half-life of 9 million years, and can date events that occurred in the first 60 million years or so of Solar System history, before most ^{182}Hf decayed away. During their early history, rocky planets differentiate into iron-rich metal cores and silicate-rich mantles. Tungsten is siderophile (it likes to bond with iron) and so partitions into the iron-rich cores. Hafnium remains in silicate and oxide minerals (it is lithophile) in the newly formed mantles. Hence, the age of core formation of a planet is recorded in the tungsten isotopic compositions of planetary materials. Core formation is thought to occur at or near the time that planets reach their final mass.

The tungsten isotope compositions of Martian meteorites have been accurately determined. But calculating the age of Mars's core formation also depends on knowing its bulk silicate Hf/W ratio. These meteorites are igneous rocks that were produced by the melting of rock deep within Mars, and that subsequently migrated and cooled near or at its surface. This migration probably resulted in fractionation of Hf and W in the magmas relative to their sources. To better determine the Hf/W ratio of bulk silicate Mars, Dauphas and Pourmand[3] used the fact that the ratio of thorium to tungsten (Th/W) in Martian meteorites is constant, and recognized that the Th/Hf ratio of Mars should not differ from the average bulk Solar System value because of the similar chemical behaviours of Th and Hf in Mars during igneous processing.

Armed with this information, the authors[3] accurately determined the Th/Hf ratio of stony meteorites (chondrites), which represent the average bulk Solar System ratio, and used this as a proxy for the Th/Hf ratio of Mars, from which they calculated its bulk silicate Hf/W ratio. By combining their calculated bulk silicate Mars Hf/W ratio with the W isotopic compositions of Martian meteorites, the authors were able to determine an age of core formation for the planet — a maximum of around 2 million to 4 million years after the Solar System began to form. This rapid formation time explains why Mars is



TAKE 27/SPL

Figure 1 | Planetary accretion. This illustration shows small rocky bodies accreting to a larger body, a protoplanet. Such accretion is thought to be the way in which protoplanets grow to become planets.

much smaller than Earth and Venus, which took tens of millions of years to form[1].

Mars has about 11% of the mass of Earth. It has a diameter of 6,800 kilometres and fits within the size distribution of rocky bodies called oligarchs[4]. Oligarchs were formed during a period of runaway growth that occurred by the accretion of up to hundreds of kilometre-sized objects known as planetesimals, and then proceeded to collide with each other to form the planets we have today[1] (Fig. 1). The authors' finding[3] that rocky bodies the size of Mars accreted within 2 million to 4 million years has ramifications for models of early planetary history.

First, oligarchs largely formed during the time when the short-lived aluminium radionuclide ^{26}Al was active (the first 2.5 million years). Aluminium concentrations in rocky bodies are typically a few per cent, and ^{26}Al can provide heat from decay. Thus, this nuclide could have provided enough heat to melt the interiors of oligarchs such that these bodies had already differentiated into core, mantle and crust before they collided with each other to form large planets. This could explain why the samarium (Sm) and neodymium (Nd) isotopic compositions of rocks from Earth and Mars indicate that these planets do not have bulk Solar System values[5–8]. During the final growth stage of planets, the collisions between Mars- to Moon-sized oligarchs were very energetic and resulted in the preferential loss of their outer shells. If these shells were made of crust formed by melting of the oligarchs' mantles during earlier differentiation, they probably contained lithophilic elements such as Sm and Nd that were not present at average Solar System proportions in the planets[9]. Hence, this formation time for oligarchs[3] predicts that the material that makes the planets would have been depleted in the elements that made the crusts — a prediction that fits the Sm–Nd isotopic compositions observed.

Second, bodies smaller than Mars, such as the Moon, should have formed at the same stage of planetary accretion as Mars, or even earlier. However, recent findings using W isotopes show that the Moon formed much later than this — as much as 60 million to 100 million years after the Solar System began to form[10]. This later time of formation strongly supports the hypothesis that the Moon formed by accretion of molten and vaporized ejecta that were produced by a collision between proto-Earth and a Mars-sized impactor very late in the formation history of the Solar System.

Third, questions remain about the accretion times of the Solar System's planets. The W isotope age calculation assumes that W and Hf were in complete or nearly complete equilibrium between silicates and iron metal during core formation in Mars as it accreted from planetesimals. If much less equilibration had occurred, then the age calculated from the W isotopes represents the prehistory of the materials that make up Mars, rather than Mars formation itself. If this is the case, the true accretion age of Mars could be more recent, and beyond the time of ^{26}Al decay as a heat source for differentiation. If so, it may well be that not all oligarchs were differentiated when they collided to grow into larger planets such as Earth. This hypothesis has its own set of compositional consequences for planetary evolution.

With such an early time for Mars accretion, which probably led to the formation of a global magma ocean[3], how do we explain the times for magma-ocean solidification of around 100 million years after the Solar System began to form that are obtained from measurements[7,8,11,12] of Lu (lutetium)–Hf and Sm–Nd chronometers in Martian meteorites? Magma oceans are not supposed to take that long to solidfy[13]. This suggests that, although Dauphas and Pourmand[3] have provided us with a key constraint on the early formation and evolution of our planets, we still have much to learn. ■

Alan Brandon *is in the Department of Earth and Atmospheric Sciences, University of Houston, Houston, Texas 77204, USA.*
e-mail: abrandon@uh.edu

1. Chambers, J. E. *Earth Planet. Sci. Lett.* **223,** 241–252 (2004).
2. Nimmo, F. & Kleine, T. *Icarus* **191,** 497–504 (2007).
3. Dauphas, N. & Pourmand, A. *Nature* **473,** 489–492 (2011).
4. Kokubo, E. & Ida, S. *Icarus* **131,** 171–178 (1998).
5. Boyet, M. & Carlson, R. W. *Science* **309,** 576–581 (2005).
6. Murphy, D. T., Brandon, A. D., Debaille, V., Burgess, R. & Ballentine, C. *Geochim. Cosmochim. Acta* **74,** 738–750 (2010).
7. Debaille, V., Brandon, A. D., Yin, Q. Z. & Jacobsen, B. *Nature* **450,** 525–528 (2007).
8. Caro, G., Bourdon, B., Halliday, A. N. & Quitté, G. *Nature* **452,** 336–339 (2008).
9. O'Neill, H. St.C & Palme, H. *Phil. Trans. R. Soc. A* **366,** 4205–4238 (2008).
10. Touboul, M., Kleine, T., Bourdon, B., Palme, H. & Wieler, R. *Nature* **450,** 1206–1209 (2007).
11. Debaille, V., Brandon, A. D., O'Neill, C., Yin, Q.-Z. & Jacobsen, B. *Nature Geosci.* **2,** 548–552 (2009).
12. Debaille, V., Yin, Q.-Z., Brandon, A. D. & Jacobsen, B. *Earth Planet. Sci. Lett.* **269,** 186–199 (2008).
13. Elkins-Tanton, L. T. *Earth Planet. Sci. Lett.* **271,** 181–191 (2008).

PROTEIN–PROTEIN INTERACTIONS

Pull-down for single molecules

An innovative marriage of techniques, combining the principles of common protein pull-down assays with single-molecule fluorescence microscopy, opens up new ways of visualizing cellular protein complexes. SEE ARTICLE P.484

PHILIP TINNEFELD

Single-molecule detection has become an essential part of such technologies as DNA sequencing and certain realizations of super-resolution fluorescence microscopy. On page 484 of this issue, Jain *et al.*[1] now present a short cut to studying protein–protein interactions at the single-molecule level.

Most biological processes are governed by assemblies of dynamically interacting proteins. Identifying all the physiological permutations of protein–protein interactions is a crucial step in unravelling the complex molecular relationships that are characteristic of living systems. Historically, protein–protein interactions have been studied using a technique called co-immunoprecipitation[2,3]. In this approach, a protein of interest (the bait) is captured from cell lysate using an appropriate antibody or a protein tag. When the bait protein is isolated, proteins that interact with it (the prey) are simultaneously captured. The captured complexes are purified and subsequently analysed using western blotting or mass spectrometry.

Jain *et al.*[1] make a short cut by immobilizing the protein complexes from a comparatively small number of lysed cells directly on a coverslip, which is then studied under the single-molecule fluorescence microscope (Fig. 1). After a washing step, the microscopic analysis can be carried out without the further laborious separation steps necessary for western blotting or mass spectrometry. Co-immobilized proteins in the complexes (the prey) are visualized using a fluorescent fusion protein or by immunofluorescence detection. Prey proteins are quantified simply by counting the number of fluorescent spots, each representing an individual protein complex, in the fluorescence images.

This scheme can thus save several hours of sample preparation. In regard to the actual measurements, the short time between cell lysis, pull-down and readout minimizes the uncertainties about whether *in vivo* interactions are maintained. As a result of the high sensitivity of the method, the amounts of proteins needed are greatly reduced. Potentially, the method might even be applied to single cells, thereby avoiding

much smaller than Earth and Venus, which took tens of millions of years to form[1].

Mars has about 11% of the mass of Earth. It has a diameter of 6,800 kilometres and fits within the size distribution of rocky bodies called oligarchs[4]. Oligarchs were formed during a period of runaway growth that occurred by the accretion of up to hundreds of kilometre-sized objects known as planetesimals, and then proceeded to collide with each other to form the planets we have today[1] (Fig. 1). The authors' finding[3] that rocky bodies the size of Mars accreted within 2 million to 4 million years has ramifications for models of early planetary history.

First, oligarchs largely formed during the time when the short-lived aluminium radionuclide ^{26}Al was active (the first 2.5 million years). Aluminium concentrations in rocky bodies are typically a few per cent, and ^{26}Al can provide heat from decay. Thus, this nuclide could have provided enough heat to melt the interiors of oligarchs such that these bodies had already differentiated into core, mantle and crust before they collided with each other to form large planets. This could explain why the samarium (Sm) and neodymium (Nd) isotopic compositions of rocks from Earth and Mars indicate that these planets do not have bulk Solar System values[5–8]. During the final growth stage of planets, the collisions between Mars- to Moon-sized oligarchs were very energetic and resulted in the preferential loss of their outer shells. If these shells were made of crust formed by melting of the oligarchs' mantles during earlier differentiation, they probably contained lithophilic elements such as Sm and Nd that were not present at average Solar System proportions in the planets[9]. Hence, this formation time for oligarchs[3] predicts that the material that makes the planets would have been depleted in the elements that made the crusts — a prediction that fits the Sm–Nd isotopic compositions observed.

Second, bodies smaller than Mars, such as the Moon, should have formed at the same stage of planetary accretion as Mars, or even earlier. However, recent findings using W isotopes show that the Moon formed much later than this — as much as 60 million to 100 million years after the Solar System began to form[10]. This later time of formation strongly supports the hypothesis that the Moon formed by accretion of molten and vaporized ejecta that were produced by a collision between proto-Earth and a Mars-sized impactor very late in the formation history of the Solar System.

Third, questions remain about the accretion times of the Solar System's planets. The W isotope age calculation assumes that W and Hf were in complete or nearly complete equilibrium between silicates and iron metal during core formation in Mars as it accreted from planetesimals. If much less equilibration had occurred, then the age calculated from the W isotopes represents the prehistory of the materials that make up Mars, rather than Mars formation itself. If this is the case, the true accretion age of Mars could be more recent, and beyond the time of ^{26}Al decay as a heat source for differentiation. If so, it may well be that not all oligarchs were differentiated when they collided to grow into larger planets such as Earth. This hypothesis has its own set of compositional consequences for planetary evolution.

With such an early time for Mars accretion, which probably led to the formation of a global magma ocean[3], how do we explain the times for magma-ocean solidification of around 100 million years after the Solar System began to form that are obtained from measurements[7,8,11,12] of Lu (lutetium)–Hf and Sm–Nd chronometers in Martian meteorites? Magma oceans are not supposed to take that long to solidfy[13]. This suggests that, although Dauphas and Pourmand[3] have provided us with a key constraint on the early formation and evolution of our planets, we still have much to learn. ■

Alan Brandon *is in the Department of Earth and Atmospheric Sciences, University of Houston, Houston, Texas 77204, USA.*
e-mail: abrandon@uh.edu

1. Chambers, J. E. *Earth Planet. Sci. Lett.* **223,** 241–252 (2004).
2. Nimmo, F. & Kleine, T. *Icarus* **191,** 497–504 (2007).
3. Dauphas, N. & Pourmand, A. *Nature* **473,** 489–492 (2011).
4. Kokubo, E. & Ida, S. *Icarus* **131,** 171–178 (1998).
5. Boyet, M. & Carlson, R. W. *Science* **309,** 576–581 (2005).
6. Murphy, D. T., Brandon, A. D., Debaille, V., Burgess, R. & Ballentine, C. *Geochim. Cosmochim. Acta* **74,** 738–750 (2010).
7. Debaille, V., Brandon, A. D., Yin, Q. Z. & Jacobsen, B. *Nature* **450,** 525–528 (2007).
8. Caro, G., Bourdon, B., Halliday, A. N. & Quitté, G. *Nature* **452,** 336–339 (2008).
9. O'Neill, H. St.C & Palme, H. *Phil. Trans. R. Soc. A* **366,** 4205–4238 (2008).
10. Touboul, M., Kleine, T., Bourdon, B., Palme, H. & Wieler, R. *Nature* **450,** 1206–1209 (2007).
11. Debaille, V., Brandon, A. D., O'Neill, C., Yin, Q.-Z. & Jacobsen, B. *Nature Geosci.* **2,** 548–552 (2009).
12. Debaille, V., Yin, Q.-Z., Brandon, A. D. & Jacobsen, B. *Earth Planet. Sci. Lett.* **269,** 186–199 (2008).
13. Elkins-Tanton, L. T. *Earth Planet. Sci. Lett.* **271,** 181–191 (2008).

PROTEIN–PROTEIN INTERACTIONS

Pull-down for single molecules

An innovative marriage of techniques, combining the principles of common protein pull-down assays with single-molecule fluorescence microscopy, opens up new ways of visualizing cellular protein complexes. SEE ARTICLE P.484

PHILIP TINNEFELD

Single-molecule detection has become an essential part of such technologies as DNA sequencing and certain realizations of super-resolution fluorescence microscopy. On page 484 of this issue, Jain *et al.*[1] now present a short cut to studying protein–protein interactions at the single-molecule level.

Most biological processes are governed by assemblies of dynamically interacting proteins. Identifying all the physiological permutations of protein–protein interactions is a crucial step in unravelling the complex molecular relationships that are characteristic of living systems. Historically, protein–protein interactions have been studied using a technique called co-immunoprecipitation[2,3]. In this approach, a protein of interest (the bait) is captured from cell lysate using an appropriate antibody or a protein tag. When the bait protein is isolated, proteins that interact with it (the prey) are simultaneously captured. The captured complexes are purified and subsequently analysed using western blotting or mass spectrometry.

Jain *et al.*[1] make a short cut by immobilizing the protein complexes from a comparatively small number of lysed cells directly on a coverslip, which is then studied under the single-molecule fluorescence microscope (Fig. 1). After a washing step, the microscopic analysis can be carried out without the further laborious separation steps necessary for western blotting or mass spectrometry. Co-immobilized proteins in the complexes (the prey) are visualized using a fluorescent fusion protein or by immunofluorescence detection. Prey proteins are quantified simply by counting the number of fluorescent spots, each representing an individual protein complex, in the fluorescence images.

This scheme can thus save several hours of sample preparation. In regard to the actual measurements, the short time between cell lysis, pull-down and readout minimizes the uncertainties about whether *in vivo* interactions are maintained. As a result of the high sensitivity of the method, the amounts of proteins needed are greatly reduced. Potentially, the method might even be applied to single cells, thereby avoiding

averaging over heterogeneous cell populations.

The simplicity of the technique, called SiMPull, is striking, raising the question of how the required specificity and impressive signal-to-noise ratio presented by Jain *et al.* is achieved. The adsorption of nonspecific proteins is minimized by using methoxy polyethylene glycol (mPEG) monolayers on the coverslips. Biotinylated PEG molecules, together with neutravidin, act as anchors for biotinylated antibodies directed against the bait protein (Fig. 1). The authors first validated the system by demonstrating efficient and specific immobilization for a polyhistidine-tagged variant of the yellow fluorescent protein (YFP). The signal-to-noise ratio was maintained at ten or more by adjusting lysate dilution factors.

The sample preparation conditions are also mild. Sensitive protein assemblies, such as intact membrane protein complexes (the β_2-adrenergic receptor), or even membrane patches, were successfully pulled down with similar efficiency and data quality. In addition, the authors show that potential problems arising from the expression of modified protein (for example, altered properties and increased or decreased expression levels compared with the wild-type protein) can be overcome by immunofluorescence detection of only endogenous complexes.

Does this work[1] present a new gold standard for analysing protein–protein interactions? To answer this question, the details of the method have to be considered. The bait protein is captured using a specific antibody or an affinity tag. Once immobilized, the proteins are detected using appropriate antibodies or, alternatively, the signal of a fused fluorescent reporter protein or fluorescent antibody can be recorded. Therefore the method is limited to well-known targets and the screening of new interaction partners is not feasible. Whether it can compete, for example, with label-free techniques such as mass spectrometry that enable the identification of protein–protein interactions with fewer constraints[4,5], is arguable. In this respect SiMPull can be viewed as an extension of western-blot analysis.

To fully appreciate the potential of the work by Jain *et al.*[1], however, the wealth of possible applications beyond the mere detection of a protein–protein interaction has to be considered. Once complexes are immobilized on the imaging surface, the trump cards offered by single-molecule fluorescence microscopy can be played, circumventing the static and dynamic averaging of common biomolecular assays. In an impressive series of examples, the authors demonstrate the variety of information that can be gained using SiMPull.

Proteins can be counted one by one, and protein expression levels can be quantified by comparison with a reference such as a recombinant protein. The stoichiometries of protein complexes can be determined by counting successive bleaching steps of single molecules, as is shown for YFP and tandem dimeric YFP, and other monomeric or dimeric proteins[6]. And multicolour imaging can be used to determine stoichiometries of heterogeneous protein complexes[7–9]. As an example, the authors show that the regulatory and catalytic subunits of the inactive tetrameric protein kinase A (PKA) are pulled down together, and that both domains are immobilized as a complex.

Moreover, immobilized complexes can be challenged in functional assays. Adding the activator cyclic AMP, which induces dissociation of the PKA complex, resulted in greatly reduced numbers of co-localized spots, confirming that constructs retain their properties. Increasing intracellular cAMP levels by external stimuli before performing SiMPull also yielded a reduced number of co-localized catalytic and regulatory domains, showing that subpopulations and changing protein interactions in cells can be revealed.

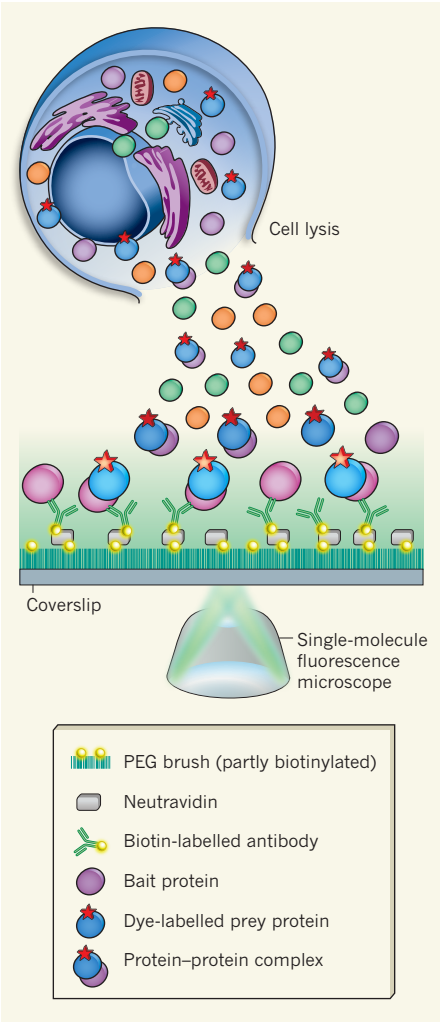


Figure 1 | The workflow of single-molecule pull-down (SiMPull)[1]. The cell lysate is applied directly to the imaging surface for single-molecule fluorescence microscopy. Protein complexes of interest are captured using specific antibodies on the surface. Prey proteins associated with the bait protein can be detected using, for example, a fluorescent dye fused to the prey. PEG, polyethylene glycol.

Above all, Jain *et al.*[1] demonstrate a sophisticated single-molecule FRET (fluorescence resonance energy transfer) experiment using immobilized PcrA helicase as a model protein. Addition of a partial duplex DNA with a 5′ overhang and ATP to the PcrA allowed the real-time observation of the helicase activity through FRET changes in the doubly labelled DNA. The direct single-molecule sample preparation by SiMPull opens up a route to study proteins in their natural complexes that are difficult to reproduce in *in vitro* experiments using recombinant proteins.

SiMPull, then, offers a great deal. It combines the principles of conventional pull-down assays with single-molecule microscopy and enables the direct visualization of cellular protein complexes. Known interactions are revealed in a robust and convenient manner, under mild preparation conditions, thereby circumventing the problems of imaging methods in living cells. SiMPull allows the determination of stoichiometries even for sensitive and short-lived protein complexes, and subpopulations arising from physiological permutations of protein–protein interactions can be revealed.

In the long term, when combined with automated workflows and microfluidics, SiMPull will possibly allow the high-throughput study of variable complex formation as a function of external stimuli such as cell stress. Clever experimental design and optimal use of the information contained in the single-molecule experiment — for example involving FRET and subnanometre localization[10] — might even allow protein pairs that physically interact, and those that happen to be in the same complex, to be distinguished. In the meantime, the numerous applications presented by Jain *et al.* will inspire other researchers, and single-molecule detection might be the key to take other important techniques to the next level. ■

Philip Tinnefeld *is at the Institute for Physical and Theoretical Chemistry, Braunschweig University of Technology, 38106 Braunschweig, Germany.*
e-mail: p.tinnefeld@tu-braunschweig.de

1. Jain, A. *et al. Nature* **473,** 484–488 (2011).
2. Puig, O. *et al. Methods* **24,** 218–229 (2001).
3. Barrios-Rodiles, M. *et al. Science* **307,** 1621–1625 (2005).
4. Vermeulen, M., Hubner, N. C. & Mann, M. *Curr. Opin. Biotechnol.* **19,** 331–337 (2008).
5. Gingras, A. C., Gstaiger, M., Raught, B. & Aebersold, R. *Nature Rev. Mol. Cell Biol.* **8,** 645–654 (2007).
6. Ulbrich, M. H. & Isacoff, E. Y. *Nature Methods* **4,** 319–321 (2007).
7. Kapanidis, A. N. *et al. Proc. Natl Acad. Sci. USA* **101,** 8936–8941 (2004).
8. Lee, J. *et al. Angew. Chem. Int. Edn* **122,** 10118–10121 (2010).
9. Stein, I. H., Steinhauer, C. & Tinnefeld, P. *J. Am. Chem. Soc.* **133,** 4193–4195 (2011).
10. Pertsinidis, A., Zhang, Y. & Chu, S. *Nature* **466,** 647–651 (2010).

PERSPECTIVE

doi:10.1038/nature10124

A 2020 vision for vaccines against HIV, tuberculosis and malaria

Rino Rappuoli[1] & Alan Aderem[2]

Acquired immune deficiency syndrome (AIDS), malaria and tuberculosis collectively cause more than five million deaths per year, but have nonetheless eluded conventional vaccine development; for this reason they represent one of the major global public health challenges as we enter the second decade of the twenty-first century. Recent trials have provided evidence that it is possible to develop vaccines that can prevent infection by human immunodeficiency virus (HIV) and malaria. Furthermore, advances in vaccinology, including novel adjuvants, prime–boost regimes and strategies for intracellular antigen presentation, have led to progress in developing a vaccine against tuberculosis. Here we discuss these advances and suggest that new tools such as systems biology and structure-based antigen design will lead to a deeper understanding of mechanisms of protection which, in turn, will lead to rational vaccine development. We also argue that new and innovative approaches to clinical trials will accelerate the availability of these vaccines.

Acquired immune deficiency syndrome (AIDS), malaria and tuberculosis are three of the most challenging infectious diseases still affecting humans (see Box 1). Since the beginning of the pandemic, AIDS has caused more than 25 million deaths, and today there are 33 million people living with HIV, 2.6 million new cases per year and 1.8 million deaths per year[1,2]. There are 225 million cases of malaria per year causing nearly one million deaths[3]. In addition, approximately one-third of the human population is infected by *Mycobacterium tuberculosis*, with 9.6 million new cases and 1.7 million deaths per year, and the bacterium becoming increasingly resistant to antibiotic therapy[4]. As one of its millennium development goals, The United Nations (UN) has elected to control and reverse the spread of these diseases by 2015. To achieve this goal the UN is relying mostly on the expanded use of therapy, education and classical measures. These include condoms in the case of AIDS and bed nets to prevent malaria[5,6].

Vaccination, which is usually the most effective intervention to control infectious diseases, was not included in the UN plan for 2015 because no vaccines are expected to be available within this period. However, new conceptual and technological advances indicate that it will be possible to develop vaccines against these diseases within the next 10 years. These advances include new prime–boost immunization regimes, new adjuvants, as well as novel methods of antigen presentation. Moreover, success will be largely dependent on our ability to use novel approaches such as systems biology to analyse data sets generated during proof-of-concept trials, leading to new insights such as the identification of correlates of protection or signatures of immunogenicity and the acceleration of large-scale clinical trials. Innovative clinical and regulatory approaches will further enhance these trials.

Systems biology and structure-based antigen design

The practice of systems biology requires capturing and integrating global sets of biological data from as many hierarchical levels as possible to visualize 'emergent properties' that are not demonstrated by their individual parts and cannot be predicted from the parts alone[7]. The response of an individual to vaccination depends on a multitude of interacting genetic, molecular and environmental factors spanning numerous temporal and spatial scales. For this reason, the tools of systems biology are particularly well suited for the analysis of vaccine studies. These data sets include molecular measurements such as DNA sequences, RNA and protein expression levels, microRNAs, protein–protein and protein–DNA interactions and metabolite biology[7,8]. These measurements are made across an array of subcellular, cellular and tissue compartments including blood, immune tissues and cellular subsets derived from them. Other relevant data include genetic variation in the populations of both people and pathogens. Finally, we need to anchor the vast array of measurements in the immune phenotypes of individuals and populations. For this reason, computation is an essential element of the systems biology approach. The inference of immunological phenotypes from global data sets spanning temporal and spatial scales exceeds the capabilities of the human mind. Computational analysis transforms thousands of data points into graphical representations that will facilitate the development of detailed computational models that directly link system phenotype to the behaviour of the protein and gene regulatory networks. Once the model is sufficiently accurate and detailed, it will allow us to predict whether novel vaccines will lead to protective responses.

One example of this approach is the systems analysis of the yellow fever vaccine YF-17D, one of the most efficacious vaccines ever developed. In this case systems biology has provided insight into its mechanism of action[9] and identified correlates of immunogenicity[10]. Expression analysis of peripheral blood mononuclear cells (PBMCs) obtained over the first 2 weeks after vaccination identified genes with expression responses or 'signatures' that are predictive of high vaccine-induced antibody and T-cell responses. The fact that these signatures were measured in peripheral blood suggests that local immune responses at the site of vaccination, which critically determine the evolution of the adaptive immunity, are reflected in this easily accessible compartment. One limitation of the study is that it could not identify a signature of efficacy because in the model used protection could not be tested. Signatures of the immune response to vaccines have further potential. For example, they can be used to predict the safety or side effects of a vaccine which would be useful in cases where the consequences of vaccination are unexpected or detrimental.

In addition to systems biology, structure-based design of novel antigens (structural vaccinology) is a powerful new tool to produce novel antigens designed to induce optimal and broadly protective immune responses[11]. Antigen design can improve vaccines by stabilizing the structure of

[1]Novartis Vaccines and Diagnostics, 53100 Siena, Italy. [2]Seattle Biomedical Research Institute, Seattle, Washington 98109, USA.

difficult antigens, by exposing and improving the immunogenicity of conserved epitopes or by engineering multiple immunodominant epitopes in one molecule to induce a broad immune response. These approaches have already been used in HIV and malaria vaccine design.

AIDS

In the early 1980s when HIV was discovered, the success of the recombinant hepatitis B virus vaccine produced in yeast led to the belief that all that was needed to make a viral vaccine was a recombinant subunit of the viral envelope. Unfortunately this has not been the case with HIV, one of the most difficult and challenging viruses discovered so far. The subunit vaccines derived from the HIV envelope were developed, tested in phase I and phase II clinical studies, and in the mid 1990s were ready to enter phase III efficacy studies; however, *in vitro* studies demonstrated that the antibodies induced by the vaccines only neutralized the virus strain used to make the vaccine and did not neutralize divergent viruses or primary viruses isolated from patients[12,13]. Therefore, phase III trials were postponed. A few years later VaxGen performed an efficacy trial using a vaccine composed of a mixture of the recombinant subunits from two clade B viruses adjuvanted with alum. This trial was performed in approximately 5,000 high-risk volunteers mainly comprising men who have sex with men[14]. A similar study with a vaccine composed of a mixture of clade B and clade E envelopes (AIDSVAX B/E) was started in Thailand on approximately 2,500 drug users[15]. The negative results of these trials were perhaps not surprising given the great antigenic diversity of the virus and the inability of the vaccines to induce antibodies able to neutralize primary isolates. The failure of the antibody-based vaccine encouraged the scientific community to focus on T-cell-mediated immunity. It had been shown that $CD8^+$ T cells against broadly conserved epitopes could be induced in non-human primates and that these were able to blunt the peak of viraemia during primary infection and maintain a low viral load for a long time after infection[12]. The enthusiasm for T-cell-based vaccines led to the design of the STEP trial, an efficacy study involving 3,000 people who were immunized either with a non-replicating adenovirus 5 (MRKAd5 HIV-1) expressing Gag/Pol/Nef or placebo. The failure of this T-cell vaccine to prevent infection or to control viral load, as had been observed in non-human primates[16], was disappointing, leading many in the field of HIV research to question the feasibility of an HIV vaccine[17]. It was therefore encouraging when the results of the RV144 trial were reported in the autumn of 2009. This trial was based on a prime–boost regime: priming with a canarypox expressing the subtype B HIV Gag, Pro and the subtype E gp120 (ALVAC-HIV) and boosting with the alum adjuvanted mix of gp120 AIDSVAX B/E. Conducted in 16,000 heterosexuals in Thailand, this trial yielded a modest 31% prevention of HIV infection[18]. Although some researchers question whether such a low efficacy is meaningful, for many the RV144 trial has renewed the hope of developing an HIV vaccine, and attempts are now being made to plan for a trial to confirm, and perhaps improve on, the results by organizing new efficacy trials based on prime–boost regimes. If successful, these new efforts could provide licensable vaccines within this decade. In the meantime, an expanded phase II trial based on a multiclade DNA priming and adenovirus 5 boost is also being conducted by the NIH Vaccine Research Center[2].

The results of three failed and one marginally successful trial could be interpreted to mean that antibodies alone or $CD8^+$ T cells alone are not effective, and that a combination of both antibodies and T cells offers marginal protection against disease. However, the immune responses underlying this protection are likely to be extraordinarily complex and only amenable to systems analysis. A comparison of the immune networks induced by various prime–boost and conventional regimes could lead to the identification of signatures of immunogenicity and possibly, in the future, of protection. Because no protective vaccines exist for HIV it is not currently possible to define correlates of protection. Thus, at the moment, we are restricted to measuring defined end points such as specific $CD4^+$ and $CD8^+$ T cells and pathogen load, which can act as useful surrogates.

Some preliminary studies are encouraging. For instance, RNA expression profiles of whole blood before and after challenge in rhesus macaques vaccinated with replicating adenovirus type 5 expressing either HIV envelope protein, simian immunodeficiency virus (SIV) Gag, or SIV Nef, followed by an HIV gp140 boost were able to identify expression signatures that distinguish vaccinated from control animals[19]. In another prime–boost study carried out in macaques, systems analysis of RNA expression profiling of PBMCs and lymph nodes identified network signatures that predicted the magnitude of specific $CD4^+$ and $CD8^+$ T-cell responses and were associated with decreased viral load (Fig. 1 and D. E. Zak *et al.*, unpublished observations). More information may also be obtained by following up some of the clinical studies that have already been performed. For instance, a subset of infected people from the STEP trial was followed for 2 years. Analysis revealed some decrease in viral load in people that carry the HLA alleles B27, B57 and B58 that are associated with more protective $CD8^+$ responses[20]. An additional observation that is still not explained is that the people that had high titres of antiadenovirus antibodies and were not circumcised were found to have an increased risk of infection[16]. Preliminary systems analysis has demonstrated that high antibody titres are associated with decreased transcription of a number of antiviral innate immune pathways including the RIG-I pathway, the TLR pathways and the inflammasome (E. Andersen-Nissen *et al.*, unpublished data).

In humans, where experimental infection with HIV is unethical, correlates of immunity may be revealed by comparing controller and progressor

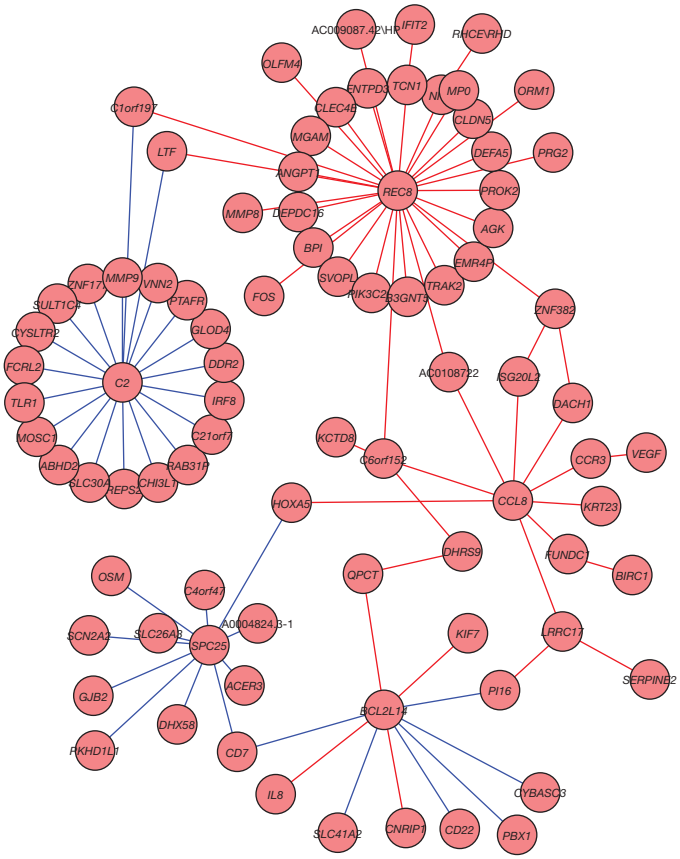


Figure 1 | Signatures that predict T-helper-cell responses after vaccination and viral load after infection. We propose that networks have stronger predictive power than do individual molecules. The network represents innate immune signatures, measured days after primary vaccination, which predict enhanced SIV Gag-specific $CD4^+$ T-cell responses, and reduced SIV load after challenge, measured months later. In this network, the circles or 'nodes' represent genes expressed in the PBMCs of macaques 6 days after vaccination. The lines between the nodes ('edges') represent associations between them. The edges indicate which combination of genes is predictive of Gag-specific $CD4^+$ T cells (blue edges) or SIV load (red edges).

populations of infected individuals (Fig. 2). HIV controllers are HIV-infected people who control viral levels sufficiently that they never progress to AIDS. A subset of these individuals control HIV through immune mechanisms. Thus, analysis of what differs between the HIV-specific immune responses of this subset and those of HIV progressors may shed light on the particular immune responses that must be elicited in the general population for an HIV vaccine to be effective. Whereas many protective genetic variations have been identified in controller populations (particularly in the major histocompatibility complex), functional differences in their HIV-specific CD8 T cells, for example, have also been identified. Deciphering the molecular networks that control these functional differences will be useful for rational vaccine design.

In the case of HIV, structure-based antigen design has been used to engineer novel gp120 molecules that are more stable, are able to expose better the universally conserved CD4 binding site and can capture broadly neutralizing antibodies[21]. Similar approaches have also been used to produce influenza HA molecules that induce antibodies against the conserved regions of the haemagglutinin located in the HA2 region[22]. In addition, today it is possible to use systematic approaches to map the repertoire of the human antibody response, identify the immunodominant neutralizing epitopes of different HIV variants and clades, and build the basis for engineering new envelope proteins able to provide broad protection.

Tuberculosis

In the case of *Mycobacterium tuberculosis*, a vaccine is available and still used to vaccinate newborns in countries with a high risk of tuberculosis infection. The vaccine was formulated a century ago and consists of Bacillus Calmette–Guerin (BCG), an attenuated strain of *Mycobacterium bovis*[23,24]. Although the overall efficacy of BCG is controversial, most agree that the vaccine is able to prevent disseminated disease and mortality in newborns and children. However, it is not able to prevent chronic infection nor to protect against pulmonary tuberculosis in adults. As a consequence, *M. tuberculosis* establishes a latent chronic infection that reactivates when there is a decrease in immune surveillance, for example in aged people, in individuals with genetic immune defects, and in those whose medication blunts their immune responses, such as a patients treated with antibodies against tumour necrosis factor-α. Immune suppression caused by HIV has become an extremely important factor in the reactivation of tuberculosis[25], and in the 15 million people co-infected by HIV and tuberculosis it is the major cause of mortality in this population[26]. Altogether, approximately two billion people carry a latent tuberculosis infection and approximately 10% will progress to active disease at some time. There are 12 vaccines against tuberculosis currently in clinical trials. Several of them are subunit vaccines consisting of recombinant antigens such as the Mtb72F fusion protein or the Ag85B-ESAT-6 fusion protein delivered with the adjuvant AS02, the Ag85-TB10.4 fusion protein delivered with the adjuvant IC31 (ref. 27), the

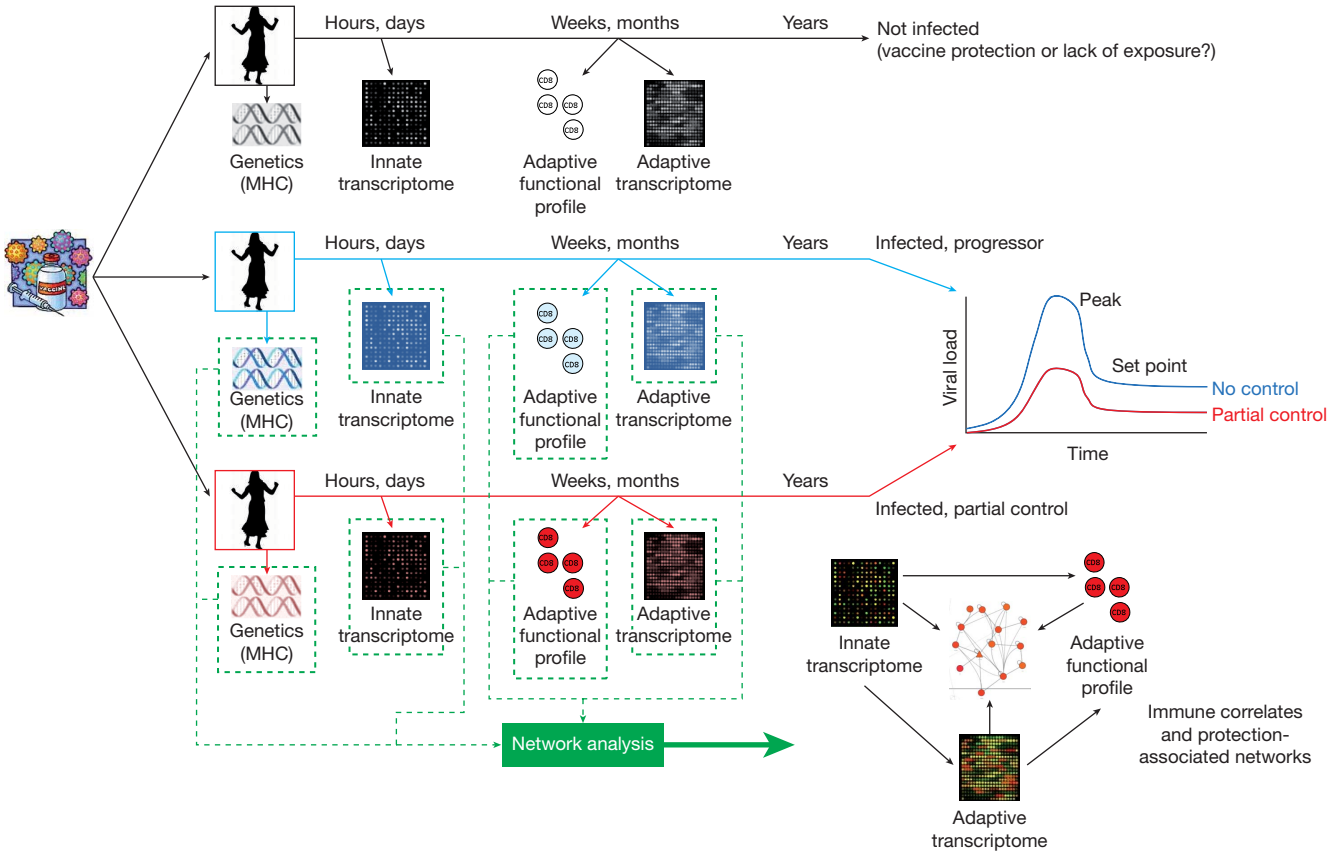


Figure 2 | Identifying novel regulatory networks associated with protection against HIV using large-scale clinical trials. Two large-scale HIV vaccine trials have shown only modest, if any, efficacy overall. Follow-up analyses are being performed to determine whether specific subgroups of participants acquire protective responses from vaccination. In principle, appropriate data mining of these trials can inform future trials by suggesting correlates of immunity and target networks that will enable rational vaccine design. For example, identification of a subpopulation of those vaccinated that became infected but controlled viral loads (red) can be compared to the population who became infected but did not control viral loads (blue). In this case, one can do systems analysis on both the innate and adaptive response to the vaccine, as well as the genetics of the individuals. Blood is collected from all of the trial participants at various times after vaccination and stored until it is known who became infected and who controlled the virus after infection. Innate cells that were collected early after vaccination and adaptive immune cells that were collected later are analysed at the molecular level using a variety of high-throughput measurements including transcriptional state, transcription-factor binding, signalling pathways, genetic variation, and metabolite and protein levels. These measurements are computationally integrated into dynamic network models that can be interpreted using visualization programs. These networks can be used as correlates of protection in subsequent vaccine trials and be used as specific molecular targets to be activated by adjuvants for rational vaccine design. This data, together with information about the individuals' major histocompatibility complex (MHC), can be used to identify better protective epitopes that can be used in future vaccines.

fusion of Ag85B-ESAT-6-Rv2660c and a variety of antigens delivered via DNA or viral vectors[25,28]. Other subunit vaccines identified by reverse vaccinology have been shown to boost BCG immunity in preclinical studies[29]. These subunit vaccines could be used to boost BCG vaccination in infants in the hope of preventing chronic infection. These vaccines could also be used in adolescents and adults to boost immunity induced by BCG or natural infection to delay or avoid reactivation. Another approach to improving tuberculosis vaccines is to re-engineer BCG to achieve better priming[30]. For example, the rBCG30 strain was engineered to overexpress antigen 85B to make it more immunogenic. Indeed, in clinical trials rBCG30 was found to induce better $CD4^+$ responses against Ag85B compared to wild-type BCG. Another engineered BCG strain was designed to engage the class I antigen presentation pathway based on the assumption that $CD8^+$ T cells are important for protection by killing tuberculosis-infected cells; this strain was therefore engineered to express the cytolysin of *Listeria monocytogenes*, a protein that enables the mycobacterium to escape from the vacuole to the cytosol, where it can be presented via class I antigen presentation pathway. The vaccine strain rBCGDUreC:Hly also has an inactivated urease gene that allows better acidification of the vacuole and improves the release of the bacterium. Preclinical studies demonstrated that this vaccine was more attenuated and more protective than BCG; it is now being tested in phase I clinical studies.

It is interesting that after a century of tuberculosis vaccine development, and after immunizing more than 3 billion people with BCG, we still know very little about immunity to *M. tuberculosis*. We still do not know why BCG induces protection, why immunity does not prevent persistent infection, what immune response would be needed to achieve sterile immunity or to prevent reactivation of latent infection. None of these questions has been answered using conventional technologies. Progress in this field will require a more comprehensive approach, such as systems biology, to test and compare different vaccines in the field and to dissect the mechanisms associated with protection. Information about immunity to tuberculosis can also be obtained by studying infected individuals. Two recent studies[31,32] used systems approaches to compare the transcripts in the blood of individuals with active infection to those of individuals who were latently infected. This investigation identified subsets of genes that correlated with the extent of the disease[31]. Although these signatures are not related to tuberculosis vaccine efficacy or immunogenicity, identification of the pathways associated with tuberculosis disease progression may help to define pathways that can be targeted in new vaccines.

Malaria

It has been known since 1967 that immunization with irradiated sporozoites can protect mice from infection with *Plasmodium berghei*[33]. It was subsequently found that humans immunized with the bites of >1,000 irradiated sporozoite-carrying mosquitoes were 100% protected from infection when challenged within 9 weeks[34]. Natural infection in endemic areas also results in protection. This is why malaria causes very severe disease and mortality in infants, children and in naive adults, but causes only mild disease in adults living in endemic areas[35,36]. The observed immunity, however, does not last indefinitely because immune people who live abroad for a period of time become susceptible again to severe malaria when they travel back to endemic countries[37].

The immunity provided by complex antigens such as irradiated sporozoites and natural infection has been very difficult to replicate using purified antigens. The best results have been obtained using the circumsporozoite protein, the most abundant antigen on the surface of the sporozoites. This protein is known to induce antibodies that inhibit the invasion of hepatocytes by sporozoites and to induce T-cell responses capable of killing infected liver cells. The antigen was expressed in a viral-like particle known as RTS,S[38]. The particle comprises 189 amino acids of the circumsporozoite antigen containing the repeat and terminal region fused to the 226 amino acids of the hepatitis B surface antigen (RTS) and the non-fused hepatitis B surface antigen.

Because the immunogenicity of RTS,S was found to be better than any recombinant circumsporozoite antigen previously expressed, it was mixed with different adjuvants and eventually used to immunize adult volunteers that were then challenged with infected mosquito bites. Surprisingly, of the three groups immunized with the RTS,S antigen, the groups receiving vaccines adjuvanted with alum plus monophosphoryl lipid A (MPL) (AS04)[39] or with the oil in water emulsion AS03 were not protected, whereas the group receiving the vaccine adjuvanted with the oil in water emulsion plus MPL and QS21 (AS02) were 86% protected from infection[40]. Interestingly, no relevant differences in antibody titres or T-cell immunity were observed between the protected group and the non-protected groups, indicating that the quality rather than the quantity of B and T cells was the key for protection. Unfortunately, at the time of this challenge study systems biology approaches were not yet available and the tools to evaluate the quality of the immune responses were limited, so that the development of the RTS,S vaccine continued empirically, without knowing why it had been so efficacious.

The vaccine was therefore tested in several clinical trials in adults and infants where it showed short-term efficacy in preventing infection ranging from 34% to 66%, and protection of 30% against clinical malaria[41,42]. The vaccine was then reformulated with a different adjuvant containing liposomes plus MPL and QS21 (AS01) and tested for efficacy; it induced short-term protection of 56% (ref. 43) during the first 8 months that decreased to 45% at 15 months[44]. Subjects are currently being enrolled for phase III efficacy trials that are expected to provide data for registration of the vaccine for use in infants and children within the next 4 years (see http://www.gavialliance.org/resources/RTS_S_fact_sheet_Oct15_FINAL_version_3.pdf). Several other approaches to malaria vaccine development have included recombinant antigens from the merozoite and gametocyte forms of the parasite, DNA- and vector-based vaccines, and irradiated sporozoites[45,46].

In light of this limited success, a comprehensive approach based on reverse vaccinology to search the genome for the best protective antigens may be necessary to develop a multicomponent vaccine that is able to confer full and long-lasting protection. Systems biology is an ideal approach to look for non-obvious differences between protective and non-protective immunity. The availability of a well-validated human challenge model where complex vaccines based on sporozoites will induce protection only if the sporozoites are irradiated but still alive, and the availability of a simple vaccine like RTS,S that can only induce protection when combined with a particular adjuvant such as AS02, represent a unique resource to look for network signatures that may distinguish a protective immune response from a non-protective one. Systems approaches will also be critical in deconvoluting the additional complexity conferred by the heterogeneity of the parasite.

In addition, structure-based design of antigens may also help to overcome the antigenic diversity of the parasite. For instance, the apical membrane antigen (AMA-1) is one of the top vaccine candidates because it can effectively inhibit the invasion of merozoites into red blood cells. However, this antigen has a low priority in vaccines considered for advanced clinical development because its antigenic diversity compromises vaccine efficacy. However, it has been shown that chimaeras of two different antigens induce inhibition of two malaria strains[47]. In this case a comprehensive approach to map the immunodominant epitopes in different variants may help the design of novel molecules able to elicit a broad immune response. Additional efforts to broaden the response to AMA1 have been recently reported[48,49]. The ability to engineer successfully antigens able to induce broad immune responses using structure-based design of immunodominant epitopes has been shown in a recent work where the meningococcus antigen factor H binding protein, which is present in three different variants, was engineered to induce protective antibodies against all natural variants of the antigen[50].

Strengths and weaknesses of systems vaccinology

Beyond signatures

Systems biology could well enable rational vaccine design. This can be achieved when we are able to define the molecular networks that control the character and quality of specific immune responses. An integration

of these molecular pathways with the correlates of protection will identify the specific networks within immune cells that need to be activated to achieve vaccine efficacy. These networks can then be selectively modulated by appropriately engineering antigens, adjuvants and vectors. Engineered vaccines can be optimized in an iterative process through a series of small-scale phase I clinical trials involving varied vaccine formulations that modulate the specific pathways suggested by the network analyses. These trials are smaller in scale because the capacity of the vaccine to activate specific networks associated with validated signatures predictive of immunogenicity is being tested as opposed to vaccine efficacy. Although the frequency of disease is rare, these signatures occur in most vaccinated people and therefore there is no need to involve thousands of participants to ensure a statistically significant number of cases. The different vaccine formulations can be evaluated based on the extent to which the protection-associated networks are triggered. These optimized vaccines can then be advanced to the next round of large efficacy trials. The ability to predict the efficacy of a vaccine early on in a trial will save both resources and lives.

Criticisms

Systems biology has often been criticized as being overly reliant on computation and there are those who suggest that computers will never be able to make biological sense of the mountains of data that are generated by the high-throughput technologies. Much of this criticism is predicated on a misunderstanding of the role of computers in systems biology. Computers are not expected to come up with biological insights *ab initio*, rather they facilitate an integration of discovery science with hypothesis-driven science to yield a holistic description of a biological system. It is also difficult to evaluate properly the success or failure of computational tools in vaccinology because, until now, trials have not been designed with systems biology analysis in mind. Another criticism of computational approaches has been the fact that most of the signatures have not yielded mechanistic insights. This has not been the intention—the goal has been for signatures that are predictive of protection which can ultimately lead to expedited vaccine trials. If these data also provide insight into the mechanism underlying protection it would be an added bonus. It has also been said that a priori analysis of the blood is naive as circulating cells do not always represent cells primarily responsible for protection. This is true, and signatures found within lymph nodes of non-human primates have given a deeper understanding of responses to vaccination (L. J. Picker *et al.*, unpublished data). At a practical level, however, blood is the only accessible means to monitor the immunological response to vaccination in humans, and correlates of protection can only be established if protective efficacy is measured in the vaccine trial.

Innovative trial design to accelerate development

During the 30 years since the discovery of HIV only four efficacy trials have been performed, an average of one trial every 8 years. Two of them have shown that anti-gp120 antibodies alone do not work; one has shown that T cells alone do not work; and one has shown that a prime–boost regime involving B and T cells may work. Altogether, only three hypotheses have been tested. Similarly, in the case of malaria, although the field has been able to benefit from experimental human challenge models and many vaccines have been tested in phase I studies[51], only two hypotheses have been tested in field efficacy trials: peptide-based vaccines and RTS,S-based vaccines. Remarkably, no efficacy trials have been performed yet for a new preventive vaccine against tuberculosis. The sequential approach, testing one hypothesis every 8 years as we have done so far, is a procedure that we cannot afford if we want to have an impact on disease in a reasonable timeframe. Accelerated clinical development can be achieved by performing more efficacy trials and by improving their design using system biology approaches to test several hypotheses in parallel and having an adaptive design[52–54] to expand the arms that are most promising (Fig. 3).

To perform more efficacy trials we need the capacity in place in those areas where diseases are prevalent as well as an adequate budget. Efficacy trials usually require budgets close to or above 100 million US dollars during the three to five years of the trial. The scientific community is often reluctant to spend this budget. However, we should keep in mind that in the case of HIV, this is less than 10% of the annual budget spent on HIV research and development. Given that the information that is obtained from a well-designed efficacy trial is of fundamental importance (even when the trial fails to show efficacy), we believe that high priority should be given to efficacy trials.

The design of efficacy trials can be improved by testing several hypotheses in parallel. For instance, several types of priming regimes paired with various boosts could be started concurrently in a large phase II study where subsets of the enrolled people are carefully monitored by systems biology approaches to test both safety and immune responses. Vaccines that elicit qualitatively similar or different immune responses can be identified, allowing more rapid discrimination of different vaccine platforms and allowing diverse concepts to be explored. The information collected during the early phases of the trial could be used to select the best arms of the trial that could be expanded to reach the statistical power to show the efficacy required for vaccine registration. Although this approach may require larger budgets during the initial phases, overall it will save money and time and will increase the probability of success. The ability to use early signatures to predict immune responses later on and therefore make early decisions on clinical trials has been recently shown to be possible. In one case, signatures in PBMCs taken 3 and 7 days

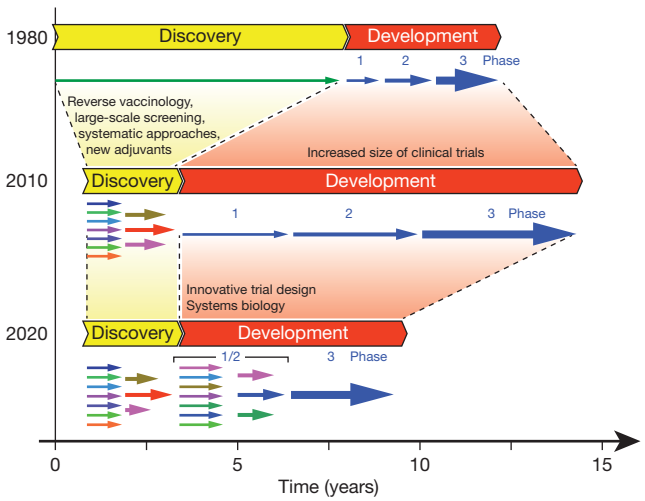
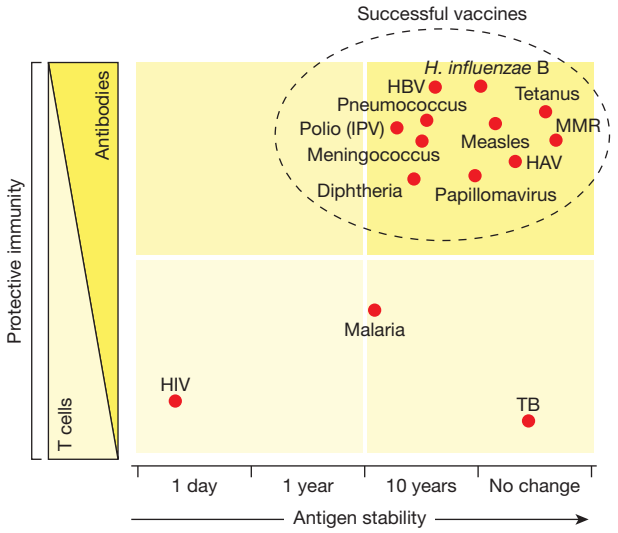


Figure 3 | Evolution of vaccine development during the last 30 years. In the 1980s discovery was the major obstacle to vaccine development. The limited technologies available allowed the development of killed, live-attenuated, toxoid, or polysaccharide vaccines only. Development used to be fast with registration requiring a few hundred subjects in clinical trials. During the 1990s, recombinant DNA technologies, conjugation and the availability of the genome sequences allowed the acceleration of the discovery phase and subsequently the discovery of vaccines against diseases that were previously impossible. High-throughput methods allowed parallel testing of multiple approaches, markedly shortening the identification of the best candidate vaccines and formulations. However, development timelines and budgets have expanded. The number of subjects required today to test safety and immunogenicity in phase I clinical trials, and in phase II studies required to define the most promising vaccine candidate, is larger than the number required to license a vaccine in 1980. Furthermore, the number of subjects required to establish safety and efficacy in phase III field registration trials has grown beyond reasonable proportions and today the licensing of a new vaccine may require up to 80,000 people in clinical trials. If phase I, II and III studies are performed sequentially, any new vaccine starting phase I requires 10 years for clinical testing. The bottom section of the figure illustrates our expectation that systems biology and adaptive design of clinical trials will accelerate the vaccine development timelines. Multiple phase I/II clinical trials can be started in parallel and be intensively and systematically monitored by systems biology until the most promising candidate emerges. At this point the arm of the most promising candidate is expanded into a phase III registration trial, saving time compared to the sequential approach. Adapted from ref. 58.

BOX 1

Challenging infectious diseases

Historically successful vaccines have been developed mostly against those pathogens that can be treated by antibodies and have a stable antigen repertoire (Box 1 Figure). HIV, malaria and tuberculosis vaccines do not fall within the cluster of successful vaccines in the graph, because of antigenic variability and the requirement of T-cell immunity for protection. Developing vaccines against these pathogens requires novel approaches.



The human immunodeficiency virus (HIV) causes the acquired immune deficiency syndrome (AIDS). The genomic sequence of the virus is highly variable. The global population of HIV viruses is divided into four major clades (A, B, C and E) that are mostly present in Africa, North America and Europe, Asia, and Africa, respectively. Within each clade the sequence variability is still huge and the virus continues to evolve and mutate within each infected patient. Virus-neutralizing antibodies and T cells induced by infection or conventional vaccination elicit a narrow immune response that is not able to provide protection against all the variants of the virus.

Malaria is caused by the *Plasmodium* parasite that infects humans via a mosquito bite. The mosquito injects the parasite in the form of a sporozoite that rapidly migrates to the liver. After 6–7 days it is released in a different form, called a merozoite, which infects red blood cells and multiplies within them. Eventually a new form of the parasite is generated (gametocyte) that is taken up by mosquitoes again. *Plasmodium falciparum* and *Plasmodium vivax* are the main human pathogens. The different stages of the parasite have different antigenic compositions and the variability of antigens within each stage has been one of the major obstacles to vaccine development.

Tuberculosis is caused by *Mycobacterium tuberculosis*, a bacterium that infects human lungs where it enters and grows within macrophages. Immune cells surround the infected macrophages and form granulomas where the bacterium may become latent for a long time. Reactivation and disease can happen when the immune system weakens, a condition that is typical of HIV. Although a live-attenuated vaccine named Bacillus Calmette–Guerin (BCG), developed almost a century ago, is used in most countries, the vaccine is not able to prevent latent infection. Therefore the reactivation of tuberculosis is still a major problem and this is exacerbated by the HIV epidemic. Although there is no variability in the antigenic repertoire, protection from infection and disease is thought to be mediated mostly by T-cell immunity, and efforts to develop vaccines that are better than BCG using conventional technologies have been unsuccessful.

after yellow fever vaccination were able to predict B- and T-cell responses measured at a later time[9,10]. In a second case, it was shown that that the frequency of $CD4^+$ T cells at day 21 after vaccination with avian influenza vaccine was able to predict the frequency of memory B cells, the presence of protective neutralizing antibodies, and the frequency of memory $CD4^+$ T cells 180 and 360 days after vaccination[55].

A change in the regulatory environment could also substantially accelerate vaccine availability. When robust correlates of protection become available by classical methods or by complex systems biology approaches, they could be used to accelerate efficacy trials and, ultimately, the implementation of the vaccine. Two examples have shown the value of accelerated implementation. The first case is the meningococcus C vaccine. A vaccine that had been tested for safety and for its ability to induce bactericidal antibodies known to correlate with protection was introduced for mass vaccination in the United Kingdom in 1999. Within 1 year from the start of vaccination the disease had disappeared from the country with a huge impact on lives saved[56]. A classical approach with an efficacy trial was going to take at least another 5 years to make the vaccine available to the general population. Similar results were obtained during the meningococcus B epidemic in New Zealand. In this case, as soon as the vaccine had been shown in phase II studies to be safe and to induce bactericidal antibodies against the strain causing the epidemic, a provisional license was issued by the regulatory agency and a large-scale, countrywide immunization campaign was started. The impact of vaccination was huge and in 1 year the epidemic disappeared[57], showing that when good correlates of protection are in place, a vaccine can be developed in 4 years without compromising safety.

As shown in Fig. 3, vaccine development can be accelerated by testing more vaccines and more vaccination regimens in parallel, and by an adaptive design of clinical trials that allows advancement to phase III registration trials without starting all over.

Conclusions

Marked progress has been made in the development of novel vaccines against the three most challenging infectious diseases of this century. Progress came from innovative vaccination concepts mediated by complex immunological mechanisms that we do not fully understand. Innovative design of clinical trials, testing several vaccines or vaccination regimes in parallel, and getting early information using systems biology approaches should allow the rapid testing of novel adjuvants, novel regimes of immunization and novel antigens. This should accelerate vaccine development and increase our understanding of the human immune system.

1. Joint United Nations Programme on HIV/AIDS. UNAIDS report on the global AIDS epidemic. ⟨http://www.unaids.org/globalreport⟩ (2010).
2. McElrath, M. J. & Haynes, B. F. Induction of immunity to human immunodeficiency virus type-1 by vaccination. *Immunity* **33,** 542–554 (2010).
3. World Health Organization. World Malaria Report 2010. ⟨http://www.who.int/malaria/world_malaria_report_2010/en/⟩ (2010).
4. World Health Organization. Global Tuberculosis Control 2010. ⟨http://www.who.int/tb/publications/global_report/en/⟩ (2010).
5. United Nations. The Millenium Development Goals Report. ⟨http://www.un.org/millenniumgoals/⟩ (2010).
6. World Health Organization. MDG 6: combat HIV/AIDS, malaria and other diseases. ⟨http://www.who.int/topics/millennium_development_goals/diseases/en/index.html⟩ (2010).
7. Zak, D. E. & Aderem, A. Systems biology of innate immunity. *Immunol. Rev.* **227,** 264–282 (2009).
   This is a comprehensive introduction to systems biology and how it can be used to study a complex biological property such as innate immunity.
8. Pulendran, B., Li, S. & Nakaya, H. I. Systems vaccinology. *Immunity* **33,** 516–529 (2010).
9. Gaucher, D. *et al.* Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses. *J. Exp. Med.* **205,** 3119–3131 (2008).
10. Querec, T. D. *et al.* Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nature Immunol.* **10,** 116–125 (2009).
    The work describes microarray analysis of PBMCs of subjects vaccinated with yellow fever vaccine and the discovery of signatures that correlate with T- and B-cell responses.

11. Dormitzer, P. R., Ulmer, J. B. & Rappuoli, R. Structure-based antigen design: a strategy for next generation vaccines. *Trends Biotechnol.* **26,** 659–667 (2008).
12. Johnston, M. I. & Fauci, A. S. An HIV vaccine—evolving concepts. *N. Engl. J. Med.* **356,** 2073–2081 (2007).
13. Mascola, J. R. & Montefiori, D. C. The role of antibodies in HIV vaccines. *Annu. Rev. Immunol.* **28,** 413–444 (2010).
14. The rgp120 HIV Vaccine Study Group. Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *J. Infect. Dis.* **191,** 654–665 (2005).
15. Pitisuttithum, P. *et al.* Randomized, double-blind, placebo-controlled efficacy trial of a bivalent recombinant glycoprotein 120 HIV-1 vaccine among injection drug users in Bangkok, Thailand. *J. Infect. Dis.* **194,** 1661–1671 (2006).
16. Buchbinder, S. P. *et al.* Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial. *Lancet* **372,** 1881–1893 (2008).
17. Burton, D. R. *et al.* Public health. A sound rationale needed for phase III HIV-1 vaccine trials. *Science* **303,** 316 (2004).
18. Rerks-Ngarm, S. *et al.* Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *N. Engl. J. Med.* **361,** 2209–2220 (2009).
A report of the results of the RV144 efficacy trial in Thailand, showing that a vaccination regime consisting of priming with a live viral vector and boosting with a recombinant protein induces a modest 31% protection from infection.
19. Palermo, R. E. *et al.* Genomic analysis reveals pre- and post challenge differences in a rhesus macaque AIDS vaccine trial: insights into mechanisms of vaccine efficacy. *J. Virol.* **85,** 1099–1116 (2011).
20. Fitzgerald, D. W. *et al.* Step Study Protocol Team. An Ad5-vectored HIV-1 vaccine elicits cell-mediated immunity but does not affect disease progression in HIV-1-infected male subjects: results from a randomized placebo-controlled trial (the Step study). *J. Infect. Dis.* **203,** 765–772 (2011).
21. Wu, X. *et al.* Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science* **329,** 856–861 (2010).
22. Wei, C.-J. *et al.* Induction of broadly neutralizing H1N1 influenza antibodies by vaccination. *Science* **329,** 1060–1064 (2010).
23. Calmette, A. *et al. La Vaccination Préventive contre la Tuberculose par le "BCG"* (Masson, 1927).
24. Behr, M. A. *et al.* Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* **284,** 1520–1523 (1999).
25. Skeiky, Y. A. W. & Sadoff, J. C. Advances in tuberculosis vaccine strategies. *Nature Rev. Microbiol.* **4,** 469–476 (2006).
26. Kaufmann, S. H. Future vaccination strategies against tuberculosis: thinking outside the box. *Immunity* **33,** 567–577 (2010).
27. Aagaard, C. *et al.* A multistage tuberculosis vaccine that confers efficient protection before and after exposure. *Nature Med.* **17,** 189–194 (2011).
28. Kaufmann, S. H., Hussey, G. & Lambert, P. H. New vaccines for tuberculosis. *Lancet* **375,** 2110–2119 (2010).
29. Bertholet, S. *et al.* A defined tuberculosis vaccine candidate boosts BCG and protects against multidrug-resistant *Mycobacterium tuberculosis*. *Sci. Transl. Med.* **2,** 53ra74 (2010).
30. Horwitz, M. A., Andersen, P. A. & Kaufmann, S. H. E. *New Generation Vaccines* 4th edn (Informa Healthcare, 2010).
31. Berry, M. P. *et al.* An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* **466,** 973–977 (2010).
32. Maertzdorf, J. *et al.* Human gene expression profiles of susceptibility and resistance in tuberculosis. *Genes Immun.* **12,** 15–22 (2011).
33. Nussenzweig, R. S., Vanderberg, J., Most, H. & Orton, C. Protective immunity produced by the injection of X-irradiated sporozoites of *Plasmodium berghei*. *Nature* **216,** 160–162 (1967).
34. Hoffman, S. L. *et al.* Protection of humans against malaria by immunization with radiation-attenuated *Plasmodium falciparum* sporozoites. *J. Infect. Dis.* **185,** 1155–1164 (2002).
35. Richie, T. L. & Saul, A. Progress and challenges for malaria vaccines. *Nature* **415,** 694–701 (2002).
36. Langhorne, J., Ndungu, F. M., Sponaas, A. M. & Marsh, K. Immunity to malaria: more questions than answers. *Nature Immunol.* **9,** 725–732 (2008).
37. Struik, S. S. & Riley, E. M. Does malaria suffer from lack of memory? *Immunol. Rev.* **201,** 268–290 (2004).
38. Gordon, D. M. *et al.* Safety, immunogenicity, and efficacy of a recombinantly produced *Plasmodium falciparum* circumsporozoite protein-hepatitis B surface antigen subunit vaccine. *J. Infect. Dis.* **171,** 1576–1585 (1995).
39. Garçon, N., Chomez, P. & Van Mechelen, M. GlaxoSmithKline Adjuvant Systems in vaccines: concepts, achievements and perspectives. *Expert Rev. Vaccines* **6,** 723–739 (2007).
40. Stoute, J. A. *et al.* A preliminary evaluation of a recombinant circumsporozoite protein vaccine against *Plasmodium falciparum* malaria. *N. Engl. J. Med.* **336,** 86–91 (1997).
This paper reports that only one of the three groups of volunteers vaccinated with the recombinant malaria antigen RTS,S was protected from infection after malaria challenge.
41. Aponte, J. J. *et al.* Safety of the RTS,S/AS02D candidate malaria vaccine in infants living in a highly endemic area of Mozambique: a double blind randomised controlled phase I/IIb trial. *Lancet* **370,** 1543–1551 (2007).
42. Alonso, P. L. *et al.* Efficacy of the RTS,S/AS02A vaccine against *Plasmodium falciparum* infection and disease in young African children: randomised controlled trial. *Lancet* **364,** 1411–1420 (2004).
43. Bejon, P. *et al.* Efficacy of RTS,S/AS01E vaccine against malaria in children 5 to 17 months of age. *N. Engl. J. Med.* **359,** 2521–2532 (2008).
44. Olotu, A. *et al.* Efficacy of RTS,S/AS01E malaria vaccine and exploratory analysis on anti-circumsporozoite antibody titres and protection in children aged 5–17 months in Kenya and Tanzania: a randomised controlled trial. *Lancet Infect. Dis.* **11,** 102–109 (2011).
45. Good, M. F. & Doolan, D. L. Malaria vaccine design: immunological considerations. *Immunity* **33,** 555–566 (2010).
46. Hoffman, S. L. *et al.* Development of a metabolically active, non-replicating sporozoite vaccine to prevent *Plasmodium falciparum* malaria. *Hum. Vaccin.* **6,** 97–106 (2010).
47. Dutta, S. *et al.* Structural basis of antigenic escape of a malaria vaccine candidate. *Proc. Natl Acad. Sci. USA* **104,** 12488–12493 (2007).
48. Kusi, K. A., Faber, B. W., Thomas, A. W. & Remarque, E. J. Humoral immune response to mixed PfAMA1 alleles; multivalent PfAMA1 vaccines induce broad specificity. *PLoS ONE* **4,** e8110 (2009).
49. Remarque, E. J., Faber, B. W., Kocken, C. H. & Thomas, A. W. A diversity-covering approach to immunization with *Plasmodium falciparum* apical membrane antigen 1 induces broader allelic recognition and growth inhibition responses in rabbits. *Infect. Immun.* **76,** 2660–2670 (2008).
50. Scarselli, M. *et al.* Rational design and structure of a meningococcal antigen inducing broad protective immunity. *Sci. Transl. Med.* (in press).
51. Sauerwein, R. W., Roestenberg, M. & Moorthy, V. S. Experimental human challenge infections can accelerate clinical malaria vaccine development. *Nature Rev. Immunol.* **11,** 57–64 (2011).
52. Freidlin, B. & Simon, R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clin. Cancer Res.* **11,** 7872–7878 (2005).
53. Koup, R. A., Graham, B. S. & Douek, D. C. The quest for a T cell-based immune correlate of protection against HIV: a story of trials and errors. *Nature Rev. Immunol.* **11,** 65–70 (2011).
54. Corey, L. *et al.* HIV-1 vaccines and adaptive trial designs. *Sci. Transl. Med.* **3,** 79ps13 (2011).
55. Galli, G. *et al.* Adjuvanted H5N1 vaccine induces early $CD4^+$ T cell response that predicts long-term persistence of protective antibody levels. *Proc. Natl Acad. Sci. USA* **106,** 3877–3882 (2009).
This study shows that $CD4^+$ T cells measured at day 21 after immunization with adjuvanted and normal influenza vaccine correlate with antibody levels measured 6 months and 1 year after vaccination, showing that biomarkers measured shortly after vaccination can predict long-term vaccine outcome.
56. Campbell, H., Borrow, R., Salisbury, D. & Miller, E. Meningococcal C conjugate vaccine: the experience in England and Wales. *Vaccine* **27** (Suppl. 2), B20–B29 (2009).
57. O'Hallahan, J., McNicholas, A., Galloway, Y., O'Leary, E. & Roseveare, C. Delivering a safe and effective strain-specific vaccine to control an epidemic of group B meningococcal disease. *N. Z. Med. J.* **122,** 48–59 (2009).
58. Masignani, V., Lattanzi, M. & Rappuoli, R. The value of vaccines. *Vaccine* **21,** Suppl. 2, s100–s103 (2003).

Acknowledgements The authors wish to thank C. Mallia for editorial assistance and G. Corsi for his contribution to the artwork. The authors also would like to thank D. Zak, K. Kennedy and S. Black for constructive criticism on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to R.R. (rino.rappuoli@novartis.com).

REVIEW

doi:10.1038/nature10108

Catalysis for fluorination and trifluoromethylation

Takeru Furuya[1]†*, Adam S. Kamlet[1]* & Tobias Ritter[1]

Recent advances in catalysis have made the incorporation of fluorine into complex organic molecules easier than ever before, but selective, general and practical fluorination reactions remain sought after. Fluorination of molecules often imparts desirable properties, such as metabolic and thermal stability, and fluorinated molecules are therefore frequently used as pharmaceuticals or materials. But the formation of carbon–fluorine bonds in complex molecules is a significant challenge. Here we discuss reactions to make organofluorides that have emerged within the past few years and which exemplify how to overcome some of the intricate challenges associated with fluorination.

Carbon–fluorine bonds have an integral role in pharmaceuticals[1,2], agrochemicals[3], materials[4] and tracers for positron emission tomography[5]. Fluorine uniquely affects the properties of organic molecules through strong polar interactions due to the atom's high electronegativity and small size[6]. For example, the introduction of fluorine into pharmaceuticals can make them more bioavailable, lipophilic and metabolically stable, and can increase the strength of a compound's interactions with a target protein[1]. Approximately 30% of all agrochemicals and 20% of all pharmaceuticals contain fluorine[1], including drugs such as Lipitor, Lexapro and Prozac.

Turning to materials, the polymer polytetrafluoroethylene, also known as Teflon, is perfluorinated (that is, all the hydrogen atoms have been replaced by fluorine atoms). The fluorine atoms are responsible for polytetrafluoroethylene's low coefficient of friction and hydrophobicity, which are properties that have made it invaluable as a non-stick coating for household cookware. Perfluorinated solvents are used as unique media for chemical reactions—when mixed with organic solvents or water, they form an immiscible 'fluorous phase' that can be useful for recovering catalysts or in purification procedures[7]. Finally, the non-natural isotope ^{18}F is the most commonly used positron-emitting isotope for molecular positron emission tomography (PET) imaging in oncology. Millions of PET scans using 2-[^{18}F]fluoro-2-deoxyglucose ([^{18}F]FDG) are performed every year[8].

Given the utility of fluorine, it is not a surprise that chemists have given the element special recognition. Yet, despite fluorine's importance and more than 100 years of organofluorine chemistry, carbon–fluorine bond formation is still challenging[9–12]. Conventional fluorination reactions that were developed in the early twentieth century are generally limited to very simple molecules[13]. Reliable fluorination of more complex molecules at specific positions is difficult. Arguably, even nature has not been able to develop a diverse set of fluorination reactions. Despite fluorine being the thirteenth most abundant element in the Earth's crust, only 21 biosynthesized natural molecules containing fluorine are known, compared to thousands with the heavier halogen homologues, chlorine and bromine[14,15]. In nature, chlorination and bromination reactions are often catalysed by haloperoxidase enzymes, but no fluoroperoxidase is known; this is likely to be a consequence of the high oxidation potential of fluorine. Additionally, the high solvation energy of the fluoride ion in aqueous media results in a tightly bound hydration shell of water molecules around the ion that lowers its nucleophilicity and therefore its reactivity. The first recognized natural fluorinating enzyme, 5′-fluoro-5′-deoxyadenosine synthase, probably dehydrates solvated fluoride in the active site, and thereby increases fluoride's nucleophilicity for the ensuing substitution reaction[16,17].

During the past five years, chemists have developed new methods to incorporate fluorine into organic molecules by making carbon–fluorine (C–F) and carbon–trifluoromethyl (C–CF_3) bonds on both aromatic rings and aliphatic chains. These new bond-forming reactions can be efficient means to access desired organic molecules that are not readily synthesized using traditional fluorination chemistry. In particular, the development of suitable catalysts for these reactions has beneficially influenced the progress of modern fluorination. In this Review, we present fundamental challenges of organofluorine chemistry and novel transition-metal-catalysed and organocatalysed C–F and C–CF_3 bond-forming reactions.

Challenges associated with C–F bond formation

Difficulties in C–F bond formation arise from the facts that fluorine is the most oxidizing and most electronegative element (Pauling electronegativity, 4.0), and that fluoride has a small ionic radius (1.33 Å; ref. 18). Owing to its electronegativity and anionic radius, fluoride, the most abundant form of the element on Earth, can form strong hydrogen bonds with, for example, water, alcohols, amines and amides[19], and therefore is typically only weakly nucleophilic in the presence of hydrogen-bond donors. Weakly nucleophilic fluoride limits access to C–F bonds via nucleophilic substitution reactions, which are a conventional and still common way to make C–F bonds[20]. When hydrogen-bond donors are meticulously excluded, fluoride is a better nucleophile, but also basic, which can lead to undesired side reactions.

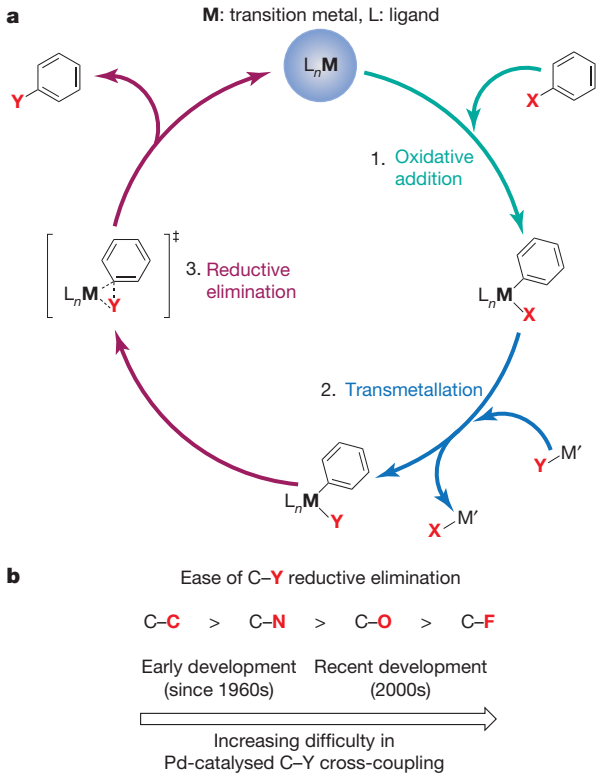
Conventional fluorination reactions that afford aryl fluorides, like the Balz-Schiemann reaction (in which anilines are converted into aryl fluorides) and the Halex process (in which halogen atoms are exchanged for fluorine atoms), generally require harsh conditions and consequently have limited substrate scopes. High temperatures or highly reactive intermediates or reagents have been the only means by which to incorporate fluorine into arenes. Reactions performed in the presence of catalysts, on the other hand, can often result in milder reaction conditions by selectively reducing the activation barriers from starting

[1]Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, USA. †Present address: SciFluor Life Sciences LLC, 33 Arch Street, Suite 3201, Boston, Massachusetts 02110, USA.
*These authors contributed equally to this work.

BOX 1

Metal-catalysed cross-coupling reactions

Metal-catalysed cross-coupling reactions are reactions that join two molecular fragments using a metal as catalyst. The 2010 Nobel Prize in Chemistry was awarded to pioneers of palladium-catalysed carbon–carbon cross-coupling reactions first disclosed over 40 years ago[96]. Since then, cross-coupling reactions have become a staple of modern organic synthesis and have been developed for virtually every element in the first and second row of the *p*-block of the periodic table[97–100]. Common examples of transition metals used in cross-coupling catalysis include palladium, copper, nickel and iron. In general, when cross-coupling reactions unite two fragments, one fragment serves as the electrophile and the other fragment serves as the nucleophile. As shown in panel a of the Box Figure below, the elementary organometallic chemistry steps of a catalysis cycle are: (1) Oxidative addition. A metal inserts into a σ-bond of the electrophile. This step increases the formal oxidation state of the metal and increases the number of ligands bound to the metal. (2) Transmetallation (ligand exchange). The nucleophile replaces a ligand on the metal. After transmetallation, both molecular fragments to be coupled are bound to the metal. (3) Reductive elimination, the actual bond-forming event that makes the organic product. Reductive elimination extrudes the new organic molecule with both molecular fragments united by a new σ-bond, leaving the metal in its original oxidation state and ready to start the catalysis cycle again. The difficulty of reductive elimination to form C–C, C–N, C–O and C–F bonds increases across the series (panel b of the Box Figure below). This trend parallels the electronegativity of the elements as well as the metal–ligand bond strength. Historically, palladium-catalysed cross-coupling reactions were developed in this order.



material to product. Catalysis has been applied to transition-metal-catalysed cross-coupling reactions with utmost success (Box 1). But until recently, fluorination reactions were notably absent from the metal-catalysed cross-coupling reaction repertoire. Carbon–fluorine bonds are strong; in fact, no other element makes stronger single bonds to carbon than fluorine does[21], and therefore, most C–F bond-forming reactions are thermodynamically feasible. A thermodynamically feasible but kinetically challenging reaction can be addressed ideally by catalysis. Conceptually, transition metal complexes have the potential to selectively reduce the barrier of activation for C–F bond formation and thus to render the thermodynamically favourable fluorination process kinetically more accessible. However, overcoming the activation barrier to C–F bond formation is challenging because metal–fluorine bonds are also strong, and thus design of appropriate catalysts is difficult.

The most challenging step in C−F bond formation via a cross-coupling approach is reductive elimination[11]; this is the step in the catalytic cycle in which both carbon and fluorine, initially bound to the metal, expel the catalyst and form a new C–F bond. For reductive elimination of two ligands to occur, there must be sufficient orbital overlap between both metal–ligand σ-bonds[12]. In general, because metal−fluorine bonds are significantly polarized towards fluorine owing to fluorine's high electronegativity and small size, electron density is lacking in the region where it is required for C−F bond formation. The high polarization of the metal–fluorine bond results in a significant ionic contribution to the bond, which strengthens it and increases the energy barrier to C–F reductive elimination. Furthermore, such reductive elimination must be faster than competing non-productive side reactions, such as hydrolysis of the metal–fluorine bond. Moreover, C–F reductive elimination is just one step in the catalysis cycle; metal–fluorine bond formation can be challenging, but is also vital to success. Methods to form the metal–fluorine bond include ligand exchange with nucleophilic fluoride and oxidative addition with electrophilic fluorination reagents. Strong, polarized and hydrolysable metal–fluorine bonds make C–F bond formation via transition metal catalysis a demanding chemical endeavour.

Metal-catalysed Ar−F bond formation

Conceptually, two fundamentally different classes of fluorination can be distinguished: nucleophilic and electrophilic fluorination. Fluoride anion (F^-) or a derivative thereof, such as tetrafluoroborate (BF_4^-), is the fluorine source in nucleophilic fluorination reactions, and an electrophilic fluorination reagent, such as XeF_2, is the source of fluorine in electrophilic fluorination reactions. Transition metals, a priori, are not biased to either nucleophilic or electrophilic fluorination, and the same metal may be successfully employed in both reaction classes. Selection of appropriate transition metals for C–F bond formation can be guided by evaluation of metal–fluorine bond strength: early transition metal fluorides generally have stronger metal–fluorine bonds compared to late transition metals owing to π-donation from the fluoride ligand into the empty *d* orbitals on the metal, and also have more polarized metal–fluorine σ-bonds. Consequently, research towards C–F bond-formation catalysis has largely focused on late transition metal complexes.

In 2002, the late transition metal copper, in the form of the electrophilic fluorination reagent CuF_2, was used in the oxidation of benzene (C_6H_6) to fluorobenzene (C_6H_5F) at 450−550 °C (ref. 22). The copper reagent can be regenerated after fluorination, and this reaction approach has the potential to lead to a practical copper-catalysed synthesis of simple fluorinated arenes. Currently, only structurally simple arenes, such as fluorobenzene, fluorotoluenes and difluorobenzenes, can be synthesized with the CuF_2-mediated process, and the reaction is characterized by low regioselectivity when substituents are present on the arene.

Regioselective functionalization of C_{aryl}−H (Ar−H) bonds by transition metals under less harsh conditions has been achieved through the use of directing groups[23]. Covalently attached to the aryl ring, directing groups coordinate to a transition metal and lower the activation energy for C−H bond cleavage preferentially, by positioning the transition metal in proximity to specific C−H bonds. The direct transformation of a C–H bond to a C–F bond is an attractive feature of directed electrophilic fluorination in terms of efficiency. Application of the directing group strategy to arene fluorination was first reported in 2006[24]. Phenylpyridine derivatives (**1**) were fluorinated at the *ortho* positions

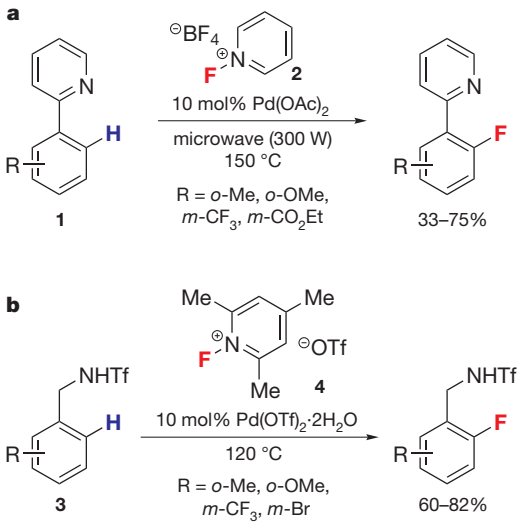


Figure 1 | Directed electrophilic palladium-catalysed Ar−F bond-forming reactions. **a**, Palladium-catalysed fluorination of organic molecules. Phenylpyridine derivatives (**1**) were fluorinated in the presence of 10 mol% of $Pd(OAc)_2$ and the electrophilic fluorination reagent *N*-fluoropyridinium tetrafluoroborate (**2**) under microwave irradiation. **b**, A palladium-catalysed directed electrophilic fluorination of C–H bonds of *N*-benzyltriflamide derivatives (**3**) with the catalyst $Pd(OTf)_2 \cdot 2H_2O$ and the electrophilic fluorination reagent *N*-fluoro-2,4,6-trimethylpyridinium triflate (**4**). Ac, acetyl; Me, methyl; Et, ethyl; Tf, trifluoromethanesulphonyl.

in the presence of $Pd(OAc)_2$ and an electrophilic fluorination reagent (Fig. 1a). A similar palladium-catalysed directed electrophilic fluorination of Ar–H bonds of *N*-benzyltriflamide derivatives (**3**) was reported in 2009 (Fig. 1b)[25]. The triflamide directing group (–NHTf, where Tf is trifluoromethanesulphonyl) can be easily converted into a variety of other functional groups. Current limitations of the directing group approach include the restriction that fluorine can only be incorporated at the position *ortho* to the directing group, the requirement for blocking groups to prevent *ortho*,*ortho'*-difluorination, and the need for a directing group itself. If the directing group is part of the desired molecule, the approach is efficient, but directing groups and functional groups that are derived from directing groups are often not desired in the final molecule, and easily removable directing groups are rare.

The mechanisms of the directed electrophilic fluorination reactions shown in Fig. 1 are still unknown. After cyclopalladation, the key C−F bond-forming event could occur either from a Pd(II) centre without change in the oxidation state of the metal (as in the electrophilic fluorination of an aryl Grignard reagent[26,27]), or from a higher oxidation state palladium complex (such as a dinuclear Pd(III)[28] or a Pd(IV)[29,30] complex) via C−F reductive elimination. Reductive elimination from transition metal complexes to form C−F bonds was long unknown. Only in 2008 was an isolated aryltransition metal fluoride complex reported to undergo C−F reductive elimination[31,32].

Transition-metal-catalysed cross coupling between an electrophile and a nucleophile is currently a more general approach for C−F bond formation, because it does not rely on directing groups. Studies of the use of palladium-, rhodium- and copper-based cross-coupling reactions for C−F bond formation have been documented since the late 1990s[33–35], but only recently has successful fluorination by catalysis been achieved, in large part owing to the development of metal complexes that can undergo C−F reductive elimination.

Theoretical studies of the fundamental difficulties associated with C−F reductive elimination from arylpalladium(II) fluoride complexes were reported in 2007[36]. Reductive elimination should occur most readily from a mononuclear, three-coordinate, 'T'-shaped palladium complex, with the aryl ligand and the fluoride ligand oriented *cis* to each other. However, 'T'-shaped arylpalladium(II) fluoride complexes are often less stable than their corresponding dimeric form, in which two 'T'-shaped palladium complexes come together, with both fluorine ligands bound to both palladium atoms. Reductive elimination from such a bis-μ-fluoride dimer is significantly more difficult than from the T-shaped monomer; in fact, to date, it has not been observed. Large ligands on palladium destabilize the dimer relative to the monomer and therefore increase the concentration of the mononuclear three-coordinate arylpalladium(II) fluoride complex for subsequent C−F reductive elimination. In line with this reasoning, the use of the bulky monodentate phosphine ligand *t*-BuXPhos resulted in C−F bond formation from an arylpalladium(II) fluoride complex, albeit in only 10% yield[36]. This was a significant and promising result, but conclusive evidence for concerted C−F reductive elimination was not obtained and other mechanisms of C−F bond formation are possible[37].

The first palladium(0)-catalysed Ar−F bond-forming cross-coupling reaction was reported in 2009 using aryl triflates (ArOTf; **5**) and CsF as a nucleophilic fluorine source (Fig. 2a)[38]. As predicted by theory, the use of a bulky monodentate phosphine ligand, *t*-BuBrettPhos (**6**)[39], to access three-coordinate arylpalladium(II) fluoride complexes was the key to success (Fig. 2b). An arylpalladium(II) fluoride complex supported by **6** was shown to be effective for C−F reductive elimination[38]. Arenes with a wide range of electronic properties and a variety of heterocycles could be fluorinated with this method. Sterically congested arenes and arenes bearing electrophilic and nucleophilic functional groups could be fluorinated as well. For a few substrates, undesired constitutional isomers were formed as by-products when *para*-electron-donating or *meta*-electron-withdrawing groups were present. Although the mechanism for the

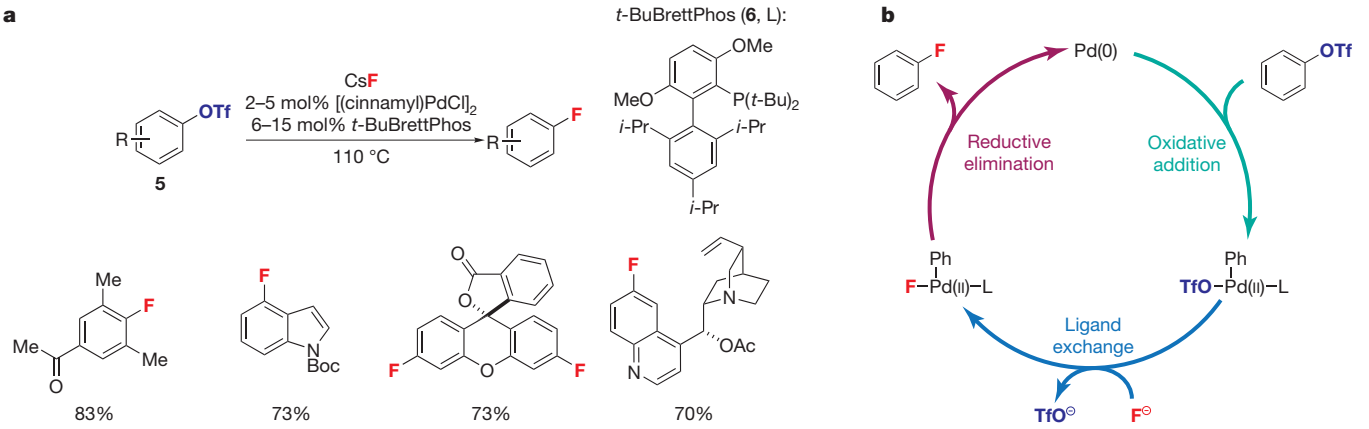


Figure 2 | Nucleophilic palladium-catalysed Ar−F bond-forming reaction. **a**, The nucleophilic palladium-catalysed Ar−F bond-forming reaction of aryl triflates (**5**), with CsF as the fluorine source, the palladium(0) catalyst precursor $[(cinnamyl)PdCl]_2$, and the sterically demanding ligand *t*-BuBrettPhos (**6**). **b**, The proposed mechanism for **a** comprises three elementary steps: oxidative addition, ligand exchange, and C−F reductive elimination. L, ligand; *t*-Bu, *tert*-butyl; *i*-Pr, *iso*-propyl; Boc, *tert*-butoxylcarbonyl; Ph, phenyl.

formation of the constitutional isomers has not yet been elucidated, the isomers could arise from a competing benzyne pathway, owing to high reaction temperatures and dried, basic fluoride. The reaction must be performed under anhydrous conditions, and substrates with protic functional groups were not demonstrated to undergo fluorination, possibly owing to the tendency of fluoride to form strong hydrogen bonds. Hydrogen-bond formation between protic functional groups or water with arylpalladium(II) fluorides could stabilize the ground state of the arylpalladium(II) fluoride complex, which increases the activation barrier to C−F reductive elimination[36]. Water could also result in hydrolysis of the Pd–F bond at a rate faster than the rate of C–F reductive elimination.

In nucleophilic fluorination, as shown in Fig. 2, fluoride serves as the nucleophile and the aryl reaction component (for example, an aryl triflate or an aryl bromide) serves as the electrophile. In 2008, C−F bond formation by a complementary approach—using a nucleophilic aryl group and an electrophilic fluorination reagent—was reported[40]. A variety of functionalized arylboronic acids are suitable substrates for transmetallation onto a palladium(II) complex; subsequent treatment with the electrophilic fluorination reagent F-TEDA-BF_4 (Selectfluor; see below) afforded the corresponding fluoroarenes. C−F bond formation occurred via fluorination of the transition metal, followed by C−F reductive elimination, which established the viability of Ar–F reductive elimination from a transition metal complex[31,32].

Reductive elimination of C−F bonds from transition metal fluorides need not be limited to palladium. The late transition metal silver has been shown to mediate the electrophilic fluorination of arylboronic acids[41] and aryl stannanes[42]. Following the initial discovery of general silver-mediated fluorination of arenes, a silver-catalysed electrophilic Ar−F bond-forming reaction for aryl stannanes (**7**) using Ag_2O and the electrophilic fluorination reagent F-TEDA-PF_6 (**8**) was developed (Fig. 3a)[43]. Several functional groups are tolerated under the reaction conditions. The reaction is applicable to late-stage fluorination of complex small molecules, including taxol (**9**), strychnine (**10**) and rifamycin (**11**) derivatives. Few nucleophilic functional groups—including certain amines and sulphides that are generally compatible with nucleophilic fluorination reactions—are incompatible with the electrophilic fluorination reaction. Current challenges associated with the silver-catalysed electrophilic fluorination include the use of toxic aryl stannane starting materials and the additional synthetic steps required for their preparation from Ar−OH or Ar−H bonds, typically via aryl triflates or aryl halides[44].

The proposed mechanism of the silver-catalysed electrophilic Ar−F bond-forming reaction consists of three elementary steps: transmetallation, silver-based oxidation by an electrophilic fluorination reagent, and C−F reductive elimination (Fig. 3b). Aryl transmetallation from tin to silver(I) affords arylsilver(I) species, which are possibly aggregated with additional silver(I) under conditions of catalysis. It was suggested that subsequent silver-based fluorination affords a multinuclear high-valent arylsilver fluoride complex, such as the dinuclear Ag^{II}–Ag^{II} complex depicted in Fig. 3b. The proposed mechanism for the silver-catalysed fluorination reaction is distinct from most conventional cross-coupling reactions owing to the redox participation of multiple metal centres. The facile C–F bond formation by silver, which enabled fluorination of complex molecules, may be due to metal–metal redox interactions that lower the barrier to C–F reductive elimination compared to mononuclear complexes[45]. Silver-catalysed carbon–heteroatom cross-coupling reactions had not been reported previously.

Metal-catalysed Ar−CF_3 bond formation

Similarly to the incorporation of fluorine, the introduction of trifluoromethyl (CF_3) groups into organic molecules can substantially alter their properties, such as metabolic stability, lipophilicity and ability to penetrate the blood–brain barrier[1–4,46,47]. Trifluoromethyl groups are distinct from other alkyl groups such as the methyl (CH_3) group, both in terms of electronic structure and reactivity; the CF_3 group has the same electronegativity as chlorine (3.2), and is similar in size to an isopropyl (*i*-Pr) group (van der Waals radius 2.2 Å)[48]. Trifluoromethyl groups, when bound to transition metals, can undergo side reactions, such as fluoride elimination[49,50], that other alkyl groups cannot. Therefore, the trifluoromethyl group should be considered more appropriately as a distinct functional group rather than as a substituted methyl group. A conventional synthesis of benzotrifluorides, arenes with a CF_3 group, involves radical chlorination of toluene derivatives followed by chlorine−fluorine exchange[51]. Only structurally simple benzotrifluorides that can tolerate such harsh reaction conditions can be accessed in this manner. Like C−F bond formation, C−CF_3 bond formation has its own challenges: the high group electronegativity of 3.2 of a trifluoromethyl group increases the activation barrier of C−CF_3 reductive elimination; only few nucleophilic and electrophilic trifluoromethylating reagents are commercially available; and the strong metal–CF_3 bonding, in part due to bonding interactions between metal *d* orbitals and the σ^*_{C-F} orbitals, make transition-metal-catalysed C–CF_3 bond formation difficult[52].

Ar−CF_3 reductive elimination from the palladium(II) complex XantphosPd(Ph)CF_3 on heating to 80 °C for 3 h was reported in 2006[53] (Xantphos is a large bidentate phosphine ligand). Whereas Ar−CF_3 reductive elimination is challenging[54–57], this result suggested that C−CF_3 bond formation by transition metal catalysis should be

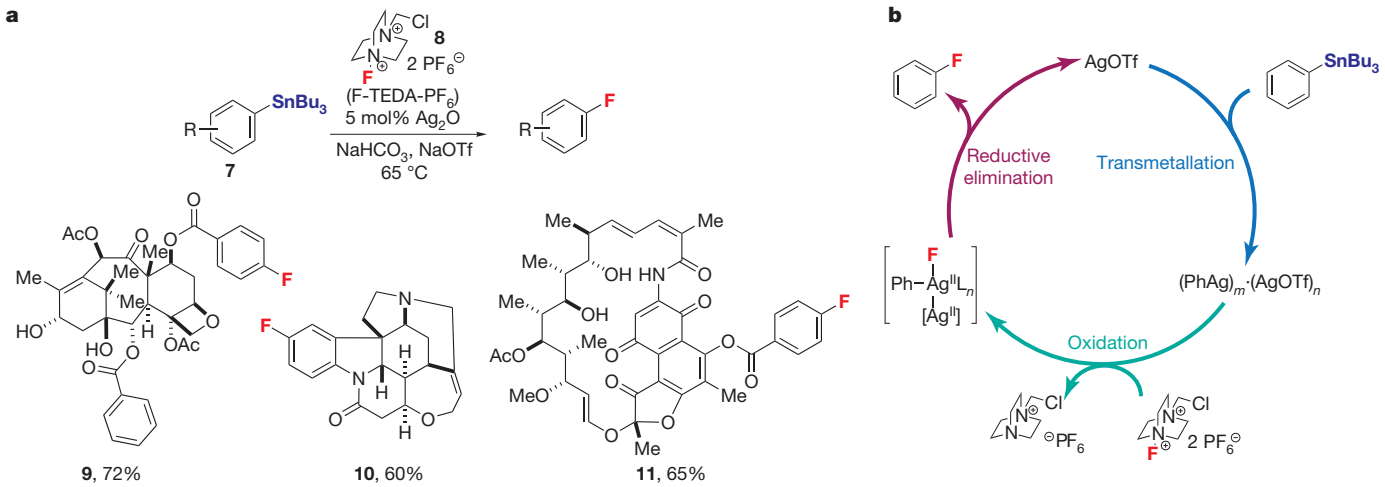


Figure 3 | Electrophilic silver-catalysed Ar−F bond-forming reaction. **a**, The silver-catalysed Ar−F bond-forming reaction. Aryl stannane derivatives (**7**) were fluorinated using 5 mol% of Ag_2O as catalyst and the electrophilic fluorination reagent F-TEDA-PF_6 (**8**). The reaction was applied to late-stage fluorination of complex small molecules, including taxol (**9**), strychnine (**10**) and rifamycin (**11**) derivatives. **b**, The proposed mechanism for **a** includes three elementary steps: transmetallation, oxidation by an electrophilic fluorination reagent, and C−F reductive elimination. Bu, butyl.

more straightforward than C–F bond formation, because Ar–F reductive elimination from the corresponding fluoride complex has not been observed. In fact, all elementary steps required for a catalysis cycle for C–CF_3 bond formation have been shown to work independently on isolated complexes[53]. The challenge for developing a palladium-catalysed aryl trifluoromethylation reaction was to develop reaction conditions that allowed all elementary steps—oxidative addition, transmetallation to make a Pd–CF_3 bond, and Ar–CF_3 reductive elimination—to proceed in the same reaction vessel, as required for catalysis.

The first Ar–CF_3 bond-forming cross-coupling reaction was reported in 1969[58]. Benzotrifluoride was obtained in 45% yield by heating iodobenzene and trifluoroiodomethane in dimethylformamide with activated copper bronze at 150 °C. Since this initial report, several modifications to the reaction conditions and reagents have been reported[59,60]. However, only in 2009 was a copper-catalysed Ar–CF_3 bond-forming reaction achieved[61]. Electron-poor aryl iodides (**12**) were converted to benzotrifluorides (**13**) with catalytic CuI and 1,10-phenanthroline (Fig. 4a). The reaction may proceed through generation of a copper-trifluoromethyl complex[62–64] followed by oxidative addition to form an arylcopper(III) intermediate[65–68], but details of the reaction mechanism remain unclear. More recently, several copper-mediated trifluoromethylation reactions have been reported[69–73].

The first palladium-catalysed Ar–CF_3 bond-forming reaction was reported in 2010 (Fig. 4b)[74]. The reaction employs aryl chlorides and (trifluoromethyl)triethylsilane ($TESCF_3$) as the CF_3 source. A large substrate scope was shown, but substrates with protic functional groups were not demonstrated to undergo trifluoromethylation, possibly because such functional groups accelerate decomposition of $TESCF_3$ or aryl(trifluoromethyl)palladium(II) and arylpalladium(II) fluoride complexes. In both the copper- and palladium-catalysed trifluoromethylation reaction, a nucleophilic trifluoromethyl unit is slowly generated *in situ* from $TESCF_3$ and KF, thus reducing the potential for side reactions to occur; the use of reagents that would generate the trifluoromethyl anion equivalent more quickly, such as (trifluoromethyl)trimethylsilane ($TMSCF_3$), result in lower trifluoromethylation yields.

Using a directing group strategy, Ar–CF_3 bond formation directly from C–H bonds can be performed with $Pd(OAc)_2$ and an electrophilic trifluoromethylation reagent (Fig. 4c)[75]. Heterocycles including pyridine, pyrimidine, imidazole and thiazole can be used as directing groups. Limitations of the reaction include the need for a directing group and the current functional group tolerance; only methoxy, chloro and methyl groups were shown to be compatible with the reaction conditions.

Catalysed C_{sp3}–F and C_{sp3}–CF_3 bond formation

Organic molecules with fluorine atoms or trifluoromethyl groups bonded to sp^3-hybridized carbon (C_{sp3}) atoms are present in pharmaceuticals, agrochemicals, dyes and materials[1–4,76–78]. Reactions using fluoride as a nucleophile for aliphatic fluorination have been known for more than 100 years[79],

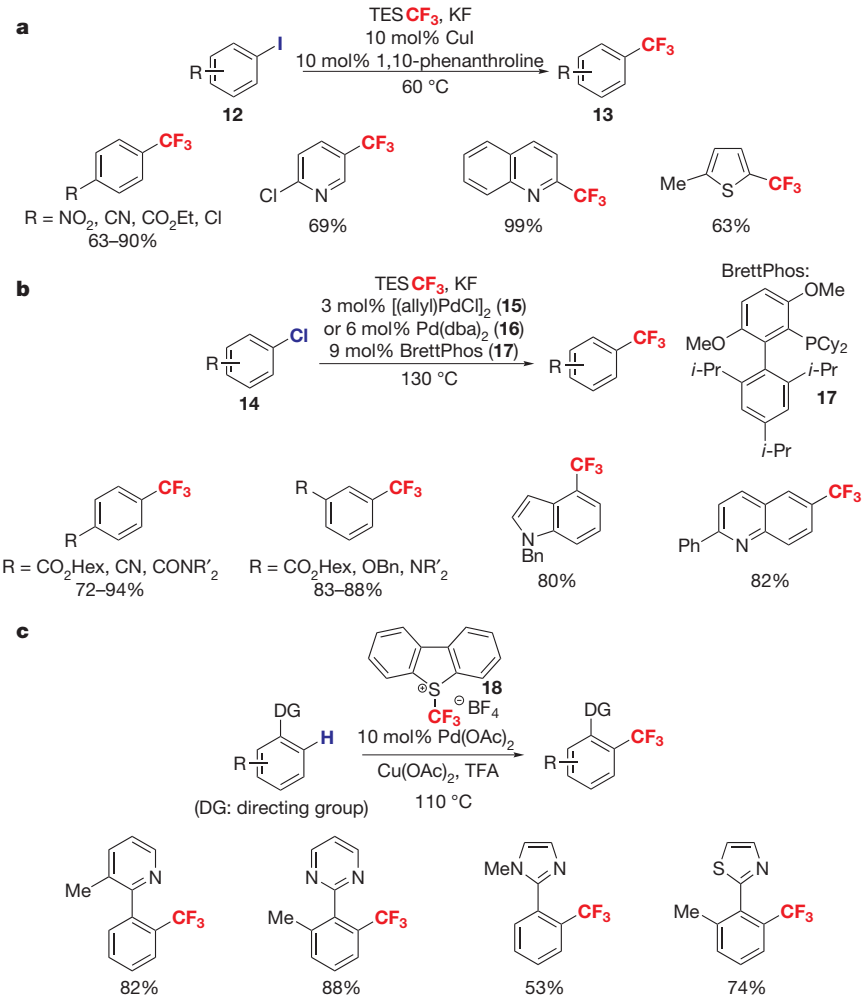


Figure 4 | Transition-metal-catalysed Ar–CF_3 bond-forming reactions. **a**, The copper-catalysed Ar–CF_3 bond-forming reaction of aryl iodides (**12**) with 10 mol% of CuI and 1,10-phenanthroline. **b**, The palladium-catalysed nucleophilic Ar–CF_3 bond-forming reaction of aryl chlorides (**14**), with $TESCF_3$ as the CF_3 source, 6 mol% of a palladium(0) precursor complex (**15** or **16**), 9 mol% of the sterically demanding ligand BrettPhos (**17**), and KF. **c**, The palladium-catalysed directed electrophilic Ar–CF_3 bond-forming reaction with 10 mol% of $Pd(OAc)_2$ and the electrophilic trifluoromethylation reagent *S*-(trifluoromethyl)dibenzothiophenium tetrafluoroborate (**18**). TES, triethylsilyl; dba, dibenzylideneacetone; Cy, cyclohexyl; Hex, hexyl; Bn, benzyl; TFA, trifluoroacetic acid.

and racemic syntheses of α-fluoro or α-trifluoromethyl carbonyl compounds were devised shortly after the development of electrophilic fluorination and trifluoromethylation reagents, respectively[10,80]. Yet, until recently, enantioselective construction of C_{sp3}−F and C_{sp3}−CF_3 bonds mainly relied on substitution reactions at existing stereogenic centres with fluoride as a nucleophile, or enantioselective addition reactions of trifluoromethyl anion equivalents to carbonyl groups[76,78] or imines[81]. In contrast to aromatic fluorination, most of the recent advances in aliphatic fluorination did not require the development of new reactivity, but rather the development of enantioselective reactions, which employed established reactivity[76–78]. Like aromatic fluorination, aliphatic fluorination benefited from developments in catalysis when compared to conventional fluorination reactions.

The electrophilic fluorination of metal enolates is a well-known process[10], but discrimination of the two enantiotopic faces of the electron-rich π-system for reaction with the electrophilic fluorinating reagent has only been achieved recently. In 2000, a titanium complex (**19**) was demonstrated to control facial selectivity in the reaction of branched α-ketoesters with Selectfluor (**20**) (Fig. 5a, top)[82]. This method was the first example of enantioselective metal-catalysed C_{sp3}−F bond formation. Two years later, an improved catalysis system was reported using a palladium catalyst (**21**) (Fig. 5a, bottom)[83]. Following these two successful examples, fluorination of the α-position of carbonyls using organometallic complexes has been investigated intensively, leading to the development of α-fluorination of malonates, α-carbamoyl esters, α-ketophosphates and α-cyanophosphates[76,77].

Similarly to the electron-rich π-systems of metal enolates, enamines can undergo electrophilic fluorination. Starting in 2005, advances in the field of organocatalytic enantioselective fluorination were reported. Fluorination reactions of cyclohexanone with Selectfluor (**20**) and proline derivatives as catalysts were investigated[84]. Immediately thereafter, three other research reports[85–87] independently disclosed enantioselective α-fluorination of aldehydes using electrophilic fluorination reagents and chiral secondary amine catalysts derived from amino acids such as **23** and **24** (Fig. 5b). The organocatalyst forms transient chiral, nucleophilic enamine intermediates, which in turn react with the electrophilic fluorinating reagent diastereoselectively. Subsequent hydrolysis of the iminium intermediate forms the chiral fluoroaldehyde and regenerates the organocatalyst. More recently, electrophilic, enantioselective fluorination of enol ethers, allyl silanes, oxindoles and cyclic ketones were reported using cinchona alkaloids as catalysts[88,89]. Nucleophilic, aliphatic fluorination by chiral organocatalysis has not yet been established, but transition metal catalysis based on chiral palladium allyl complexes can be used to make allylic fluorides[90] enantioselectively[91].

Enantioselective α-trifluoromethylation of carbonyls can proceed analogously to organocatalysed fluorination, with appropriate electrophilic trifluoromethylation reagents. Two such reactions of aldehydes have been reported. Aldehydes were trifluoromethylated enantioselectively with the hypervalent iodine reagent **27**[92] as the electrophilic reagent, and using the chiral imidazolidinone catalyst **26** (Fig. 5c, top)[93].

In contrast to the other organocatalysed reactions discussed here, organocatalysed trifluoromethylation of aldehydes via photoredox catalysis proceeds by a mechanism distinct from conventional fluorination and trifluoromethylation reactivity. Photoredox catalysis operates via one-electron pathways[94], whereas the other presented organocatalysed reactions probably proceed via two-electron pathways. Using the trifluoromethylation reagent iodotrifluoromethane, the chiral organocatalyst **28**, the iridium catalyst **29**, and light from a fluorescent light bulb, aldehydes were transformed into the corresponding α-trifluoromethyl aldehydes with high enantioselectivity (Fig. 5c, bottom)[95]. Under the reaction conditions, trifluoromethyl radicals are generated by single electron transfer from the photolytically activated Ir catalyst, and these radicals in turn oxidize *in situ* generated enamines to form C−CF_3 bonds.

Conclusions

Fluorinated organic molecules are often valuable but are generally challenging to synthesize efficiently. Fluorination reactions developed in the past five years now give access to complex fluorinated molecules that were not readily available before. The recent success in fluorine incorporation can be attributed to the design of previously unavailable transition metal complexes, and the merger of modern synthesis techniques, such as organocatalysis, with fluorination chemistry. Catalysis has played a major part in the recent development of organofluorine chemistry. For

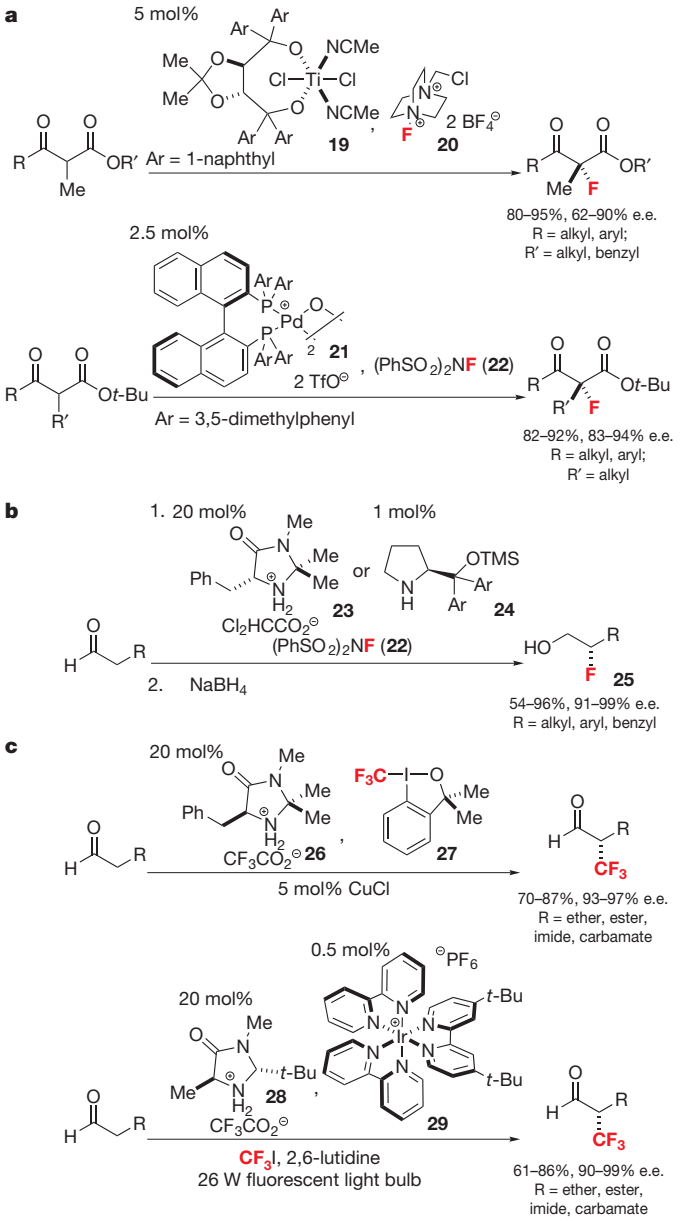


Figure 5 | Catalytic enantioselective C_{sp3}−F and C_{sp3}−CF_3 bond-forming reactions. a, Metal-catalysed enantioselective C_{sp3}−F bond-forming reactions. Branched β-ketoesters were fluorinated using 5 mol% of a Ti-TADDOL catalyst (**19**) and Selectfluor (**20**) or 2.5 mol% of μ-hydroxo-palladium-BINAP complex (**21**) and *N*-fluorobenzenesulphonimide (**22**). **b**, Examples of organocatalytic enantioselective C_{sp3}−F bond-forming reactions. Amino acid-derived organocatalysts (**23** and **24**) and *N*-fluorobenzenesulphonimide (**22**) were used to fluorinate α-unbranched aldehydes. Owing to the potentially facile racemization of α-fluoroaldehydes, the corresponding fluorohydrins (**25**) were isolated after reduction with $NaBH_4$ in 54−96% yield and 91−99% enantiomeric excess (e.e.). **c**, Enantioselective C_{sp3}−CF_3 bond-forming reactions. The mechanism of the two presented reactions differ conceptually, but both afford α-trifluoromethylated aldehydes in good yield and high enantioselectivity either using hypervalent iodine **27** as the CF_3 source and amine catalyst **26**, or using trifluoroiodomethane as the CF_3 source, 20 mol% amine catalyst **28**, 0.5 mol% Ir catalyst **29** and light (from a fluorescent household light bulb). TMS, trimethylsilyl.

example, new catalysts can selectively lower the activation barrier of C−F and C−CF_3 bond formation in aromatic fluorination and trifluoromethylation, respectively, and new chiral catalysts can distinguish between the enantiotopic faces of nucleophiles for aliphatic C–F and C–CF_3 bond formation.

Better prediction of the reactivity of well-defined transition metal complexes has supported the advances in nucleophilic aromatic fluorination. For example, rational ligand development was crucial for the palladium-catalysed C−F cross-coupling reaction described in Fig. 2. Readily available starting materials such as aryl triflates and fluoride, arguably the simplest and most desirable source of fluorine, can be employed in this reaction. Yet, whenever fluoride is used, its basicity and the basicity of the transition metal complexes derived from it are often problematic, because water and protic functional groups inhibit the desired reactivity. Future nucleophilic fluorination reactions will benefit from the availability of transition metal complexes of lower basicity, which, most probably, will be achieved through further design of transition metal complexes and ligands. Moreover, reaction methods based on other transition metals may be suitable for more widely applicable nucleophilic fluorination. In particular, the coinage metals (copper, silver and gold) have shown intriguing reactivity and merit further study.

Current electrophilic fluorination reactions have different challenges. Silver-catalysed electrophilic fluorination of aryl stannanes has the largest demonstrated substrate scope and is amenable to the fluorination of complex molecules. However, aryl stannanes are toxic and more difficult to synthesize than aryl triflates or halides, which reduces the method's practicality. Practical cross-coupling reactions should employ readily accessible starting materials, such as aryl chlorides and phenols. Ideally, regioselective electrophilic fluorination reactions would transform C−H bonds into C–F bonds directly. C−H bond functionalization reactions would be the most efficient means of incorporating fluorine into complex molecules, but are currently limited to very simple arenes, such as benzene, and arenes with directing groups. Future advances in the field of selective C−H functionalization combined with modern fluorination chemistry will probably result in practical fluorination reactions. For example, a general, regioselective fluorination of C–H bonds using fluoride and an economically viable oxidant would significantly advance the field.

The reactions presented in this Review have begun to address some of the unmet needs in organofluorine chemistry. In medicinal chemistry, milligram to gram quantities of functionalized fluorinated molecules are more readily accessible now than before. On the other hand, current methods still lack practicality and cost efficiency for general use in large-scale manufacturing. And although fluorination to prepare tracer molecules for positron emission tomography (PET) with the isotope ^{18}F only requires small amounts of material, the recent advances in fluorination technology have not given access to general ^{18}F-tracer synthesis, because the stringent reaction requirements for practical and general ^{18}F-fluorination are not met. Future research in fluorination chemistry will need to focus on the development of more general and practical fluorination reactions.

Received 15 December 2010; accepted 7 April 2011.

1. Müller, K., Faeh, C. & Diederich, F. Fluorine in pharmaceuticals: looking beyond intuition. *Science* **317,** 1881–1886 (2007).
2. Purser, S., Moore, P. R., Swallow, S. & Gouverneur, V. Fluorine in medicinal chemistry. *Chem. Soc. Rev.* **37,** 320–330 (2008).
3. Jeschke, P. The unique role of fluorine in the design of active ingredients for modern crop production. *ChemBioChem* **5,** 570–589 (2004).
4. Hung, M. H., Farnham, W. B., Feiring, A. E. & Rozen, S. in *Fluoropolymers: Synthesis* Vol. 1 (eds Hougham, G., Cassidy, P. E., Johns, K. & Davidson, T.) 51–66 (Plenum, 1999).
5. Ametamey, S. M., Honer, M. & Schubiger, P. A. Molecular imaging with PET. *Chem. Rev.* **108,** 1501–1516 (2008).
6. O'Hagan, D. Understanding organofluorine chemistry. An introduction to the C−F bond. *Chem. Soc. Rev.* **37,** 308–319 (2008).
7. Curran, D. P. Strategy-level separations in organic synthesis: from planning to practice. *Angew. Chem. Int. Edn* **37,** 1174–1196 (1998).
8. Patterson, J. C. II & Mosley, M. L. How available is positron emission tomography in the United States? *Mol. Imaging Biol.* **7,** 197–200 (2005).
9. Kirk, K. L. Fluorination in medicinal chemistry: methods, strategies, and recent developments. *Org. Process Res. Dev.* **12,** 305–321 (2008).
10. Furuya, T., Kuttruff, C. A. & Ritter, T. Carbon–fluorine bond formation. *Curr. Opin. Drug Discov. Dev.* **11,** 803–819 (2008).
11. Grushin, V. V. The organometallic fluorine chemistry of palladium and rhodium: studies toward aromatic fluorination. *Acc. Chem. Res.* **43,** 160–171 (2010).
12. Furuya, T., Klein, J. E. M. N. & Ritter, T. C–F bond formation for the synthesis of aryl fluorides. *Synthesis* 1804–1821 (2010).
13. Kirsch, P. *Modern Fluoroorganic Chemistry: Synthesis, Reactivity, Applications* (Wiley, 2004).
14. Gribble, G. W. in *Progress in the Chemistry of Organic Natural Products* Vol. 68 (eds Herz, W., Kirby, G. W., Moore, R. E., Steglich, W. & Tamm, C.) 1–498 (Springer, 1996).
15. Gribble, G. W. in *Progress in the Chemistry of Organic Natural Products* Vol. 91 (eds Kinghord, A. D., Falk, H. & Kobayashi, J.) 1–613 (Springer, 2009).
16. O'Hagan, D., Schaffrath, C., Cobb, S. L., Hamilton, J. T. G. & Murphy, C. D. Biochemistry: Biosynthesis of an organofluorine molecule. *Nature* **416,** 279 (2002).
17. Dong, C. *et al.* Crystal structure and mechanism of a bacterial fluorinating enzyme. *Nature* **427,** 561–565 (2004).
18. Shannon, R. D. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallogr. A* **32,** 751–767 (1976).
19. Emsley, J. Very strong hydrogen bonds. *Chem. Soc. Rev.* **9,** 91–124 (1980).
20. Adams, D. J. & Clark, J. H. Nucleophilic routes to selectively fluorinated aromatics. *Chem. Soc. Rev.* **28,** 225–231 (1999).
21. Luo, Y.-R. *Handbook of Bond Dissociation Energies in Organic Compounds* (CRC Press, 2002).
22. Subramanian, M. A. & Manzer, L. E. A ''greener'' synthetic route for fluoroaromatics via copper(II) fluoride. *Science* **297,** 1665 (2002).
23. Cope, A. C. & Siekman, R. W. Formation of covalent bonds from platinum or palladium to carbon by direct substitution. *J. Am. Chem. Soc.* **87,** 3272–3273 (1965).
24. Hull, K. L., Anani, W. Q. & Sanford, M. S. Palladium-catalyzed fluorination of carbon–hydrogen bonds. *J. Am. Chem. Soc.* **128,** 7134–7135 (2006).
Describes the first Pd-catalysed C−F bond formation.
25. Wang, X., Mei, T.-S. & Yu, J.-Q. Versatile $Pd(OTf)_2{\cdot}2H_2O$-catalyzed *ortho*-fluorination using NMP as a promoter. *J. Am. Chem. Soc.* **131,** 7520–7521 (2009).
26. Yamada, S., Gavryushin, A. & Knochel, P. Convenient electrophilic fluorination of functionalized aryl and heteroaryl magnesium reagents. *Angew. Chem. Int. Edn* **49,** 2215–2218 (2010).
27. Anbarasan, P., Neumann, H. & Beller, M. Efficient synthesis of aryl fluorides. *Angew. Chem. Int. Edn* **49,** 2219–2222 (2010).
28. Powers, D. C. & Ritter, T. Bimetallic Pd(III) complexes in palladium-catalysed carbon−heteroatom bond formation. *Nature Chem.* **1,** 302–309 (2009).
29. Kaspi, A. W., Yahav-Levi, A., Goldberg, I. & Vigalok, A. Xenon difluoride induced aryl iodide reductive elimination: a simple access to difluoropalladium(II) complexes. *Inorg. Chem.* **47,** 5–7 (2008).
30. Ball, N. D. & Sanford, M. S. Synthesis and reactivity of a mono-σ-aryl palladium(IV) fluoride complex. *J. Am. Chem. Soc.* **131,** 3796–3797 (2009).
31. Furuya, T. & Ritter, T. Carbon−fluorine reductive elimination from a high-valent palladium fluoride. *J. Am. Chem. Soc.* **130,** 10060–10061 (2008).
Describes the first confirmed Ar−F reductive elimination from a transition metal complex.
32. Furuya, T. *et al.* Mechanism of C−F reductive elimination from palladium(IV) fluorides. *J. Am. Chem. Soc.* **132,** 3793–3807 (2010).
33. Fraser, S. L., Antipin Yu, M., Khroustalyov, V. N. & Grushin, V. V. Molecular fluoro palladium complexes. *J. Am. Chem. Soc.* **119,** 4769–4770 (1997).
34. Pilon, M. C. & Grushin, V. V. Synthesis and characterization of organopalladium complexes containing a fluoro ligand. *Organometallics* **17,** 1774–1781 (1998).
35. Grushin, V. V. Palladium fluoride complexes: one more step toward metal-mediated C−F bond formation. *Chem. Eur. J.* **8,** 1006–1014 (2002).
36. Yandulov, D. V. & Tran, N. T. Aryl–fluoride reductive elimination from Pd(II): feasibility assessment from theory and experiment. *J. Am. Chem. Soc.* **129,** 1342–1358 (2007).
37. Grushin, V. V. & Marshall, W. J. Ar–F reductive elimination from palladium(II) revisited. *Organometallics* **26,** 4997–5002 (2007).
38. Watson, D. A. *et al.* Formation of ArF from LPdAr(F): catalytic conversion of aryl triflates to aryl fluorides. *Science* **325,** 1661–1664 (2009).
Reports the first functional-group-tolerant Pd-catalysed Ar−F bond formation using aryl triflates and fluoride.
39. Fors, B. P., Watson, D. A., Biscoe, M. R. & Buchwald, S. L. A highly active catalyst for Pd-catalyzed amination reactions: cross-coupling reactions using aryl mesylates and the highly selective monoarylation of primary amines using aryl chlorides. *J. Am. Chem. Soc.* **130,** 13552–13554 (2008).
40. Furuya, T., Kaiser, H. M. & Ritter, T. Palladium-mediated fluorination of arylboronic acids. *Angew. Chem. Int. Edn* **47,** 5993–5996 (2008).
41. Furuya, T. & Ritter, T. Fluorination of boronic acids mediated by silver(I) triflate. *Org. Lett.* **11,** 2860–2863 (2009).
42. Furuya, T., Strom, A. E. & Ritter, T. Silver-mediated fluorination of functionalized aryl stannanes. *J. Am. Chem. Soc.* **131,** 1662–1663 (2009).
43. Tang, P., Furuya, T. & Ritter, T. Silver-catalyzed late-stage fluorination. *J. Am. Chem. Soc.* **132,** 12150–12154 (2010).
Reports the first functional-group-tolerant Ag-catalysed Ar−F bond formation using aryl stannanes and an electrophilic fluorinating reagent.

44. Azizian, H., Eaborn, C. & Pidcock, A. Synthesis of organotrialkylstannanes. The reaction between organic halides and hexaalkyldistannanes in the presence of palladium complexes. *J. Organomet. Chem.* **215,** 49–58 (1981).
45. Powers, D. C., Benitez, D., Tkatchouk, E., Goddard, W. A. III & Ritter, T. Bimetallic reductive elimination from dinuclear Pd(III) complexes. *J. Am. Chem. Soc.* **132,** 14092–14103 (2010).
46. Ma, J.-A. & Cahard, D. Strategies for nucleophilic, electrophilic, and radical trifluoromethylations. *J. Fluor. Chem.* **128,** 975–996 (2007).
47. Shimizu, M. & Hiyama, T. Modern synthetic methods for fluorine-substituted target molecules. *Angew. Chem. Int. Edn* **44,** 214–231 (2005).
48. Bott, G., Field, L. D. & Sternhell, S. Steric effects. A study of a rationally designed system. *J. Am. Chem. Soc.* **102,** 5618–5626 (1980).
49. Jensen, M. B. *et al.* Reactivity and structure of CF_3I on Ru(001). *J. Phys. Chem.* **99,** 8736–8744 (1995).
50. Liu, Z.-M., Zhou, X.-L., Kiss, J. & White, J. M. Interaction of CF_3I with Pt(111). *Surf. Sci.* **286,** 233–245 (1993).
51. Yagupolskii, L. M. in *Houben-Weyl: Methods of Organic Chemistry* Vol. E10a, *Organo-Fluorine Compounds* (eds Baasner, B., Hagemann, H. & Tatlow, J.-C.) 509–534 (Thieme, 2000).
52. Clark, H. C. & Tsai, J. H. Bonding in fluorinated organometallic compounds. *J. Organomet. Chem.* **7,** 515–517 (1967).
53. Grushin, V. V. & Marshall, W. J. Facile Ar–CF_3 bond formation at Pd. Strikingly different outcomes of reductive elimination from $[(Ph_3P)_2Pd(CF_3)Ph]$ and $[(Xantphos)Pd(CF_3)Ph]$. *J. Am. Chem. Soc.* **128,** 12644–12645 (2006).
54. Culkin, D. A. & Hartwig, J. F. Carbon–carbon bond-forming reductive elimination from arylpalladium complexes containing functionalized alkyl groups. Influence of ligand steric and electronic properties on structure, stability, and reactivity. *Organometallics* **23,** 3398–3416 (2004).
55. Grushin, V. V. & Marshall, W. J. Unexpected H_2O-induced Ar–X activation with trifluoromethylpalladium(II) aryls. *J. Am. Chem. Soc.* **128,** 4632–4641 (2006).
56. Ball, N. D., Kampf, J. W. & Sanford, M. S. Aryl–CF_3 bond-forming reductive elimination from palladium(IV). *J. Am. Chem. Soc.* **132,** 2878–2879 (2010).
57. Ye, Y., Ball, N. D., Kampf, J. W. & Sanford, M. S. Oxidation of a cyclometalated Pd(II) dimer with "CF_3^+": formation and reactivity of a catalytically competent monomeric Pd(IV) aquo complex. *J. Am. Chem. Soc.* **132,** 14682–14687 (2010).
58. Mcloughlin, V. C. R. & Thrower, J. A route to fluoroalkyl-substituted aromatic compounds involving fluoroalkylcopper intermediates. *Tetrahedron* **25,** 5921–5940 (1969).
59. Kobayashi, Y. & Kumadaki, I. Trifluoromethylation of aromatic compounds. *Tetrahedr. Lett.* **10,** 4095–4096 (1969).
60. Carr, G. E., Chambers, R. D., Holmes, T. F. & Parker, D. G. Sodium perfluoroalkane carboxylates as sources of perfluoroalkyl groups. *J. Chem. Soc. Perkin Trans. I* 921–926 (1988).
61. Oishi, M., Kondo, H. & Amii, H. Aromatic trifluoromethylation catalytic in copper. *Chem. Commun.* 1909–1911 (2009).
62. Wiemers, D. A. & Burton, D. J. Pregeneration, spectroscopic detection, and chemical reactivity of (trifluoromethyl)copper, an elusive and complex species. *J. Am. Chem. Soc.* **108,** 832–834 (1986).
63. Dubinina, G. G., Furutachi, H. & Vicic, D. A. Active trifluoromethylating agents from well-defined copper(I)-CF_3 complexes. *J. Am. Chem. Soc.* **130,** 8600–8601 (2008).
64. Dubinina, G. G., Ogikubo, J. & Vicic, D. A. Structure of bis(trifluoromethyl)cuprate and its role in trifluoromethylation reactions. *Organometallics* **27,** 6233–6235 (2008).
65. Monnier, F. & Taillefer, M. Catalytic C–C, C–N, and C–O Ullmann-type coupling reactions. *Angew. Chem. Int. Edn* **48,** 6954–6971 (2009).
66. Altman, R. A., Hyde, A. M., Huang, X. & Buchwald, S. L. Orthogonal Pd- and Cu-based catalyst systems for C- and N-arylation of oxindoles. *J. Am. Chem. Soc.* **130,** 9613–9620 (2008).
67. Tye, J. W., Weng, Z., Johns, A. M., Incarvito, C. D. & Hartwig, J. F. Copper complexes of anionic nitrogen ligands in the amidation and imidation of aryl halides. *J. Am. Chem. Soc.* **130,** 9971–9983 (2008).
68. Huffman, L. M. & Stahl, S. S. Carbon–nitrogen bond formation involving well-defined aryl–copper(III) complexes. *J. Am. Chem. Soc.* **130,** 9196–9197 (2008).
69. Knauber, T., Arikan, F., Röschenthaler, G.-V. & Gooßen, L. J. Copper-catalyzed trifluoromethylation of aryl iodides with potassium (trifluoromethyl) trimethoxyborate. *Chem. Eur. J.* **17,** 2689–2697 (2011).
70. Chu, L. & Qing, F.-L. Copper-mediated oxidative trifluoromethylation of boronic acids. *Org. Lett.* **12,** 5060–5063 (2010).
71. Zhang, C.-P. *et al.* Copper-mediated trifluoromethylation of heteroaromatic compounds by trifluoromethyl sulfonium salts. *Angew. Chem. Int. Edn* **50,** 1896–1900 (2011).
72. Senecal, T. D., Parsons, A. T. & Buchwald, S. L. Room temperature aryl trifluoromethylation via copper-mediated oxidative cross-coupling. *J. Org. Chem.* **76,** 1174–1176 (2011).
73. Morimoto, H., Tsubogo, T., Litvinas, N. D. & Hartwig, J. F. A broadly applicable copper reagent for trifluoromethylations and perfluoroalkylations of aryl iodides and bromides. *Angew. Chem. Int. Edn* **50,** 3793–3798 (2011); published online 25 March 2011.
74. Cho, E. J. *et al.* The palladium-catalyzed trifluoromethylation of aryl chlorides. *Science* **328,** 1679–1681 (2010).
Describes the first functional-group-tolerant Pd-catalysed Ar–CF_3 bond formation.
75. Wang, X., Truesdale, L. & Yu, J.-Q. Pd(II)-catalyzed *ortho*-trifluoromethylation of arenes using TFA as a promoter. *J. Am. Chem. Soc.* **132,** 3648–3649 (2010).
76. Ma, J.-A. & Cahard, D. Asymmetric fluorination, trifluoromethylation, and perfluoroalkylation reactions. *Chem. Rev.* **108,** PR1–PR43 (2008).
77. Lectard, S., Hamashima, Y. & Sodeoka, M. Recent advances in catalytic enantioselective fluorination reactions. *Adv. Synth. Catal.* **352,** 2708–2732 (2010).
78. Shibata, N., Mizuta, S. & Kawai, H. Recent advances in enantioselective trifluoromethylation reactions. *Tetrahedr. Asymm.* **19,** 2633–2644 (2008).
79. Young, S. Note on the formation of an alcoholic fluoride. *J. Chem. Soc.* **39,** 489–497 (1881).
80. Umemoto, T. & Adachi, K. New method for trifluoromethylation of enolate anions and applications to regio-, diastereo- and enantioselective trifluoromethylation. *J. Org. Chem.* **59,** 5692–5699 (1994).
81. Kawai, H., Kusuda, A., Nakamura, S., Shiro, M. & Shibata, N. Catalytic enantioselective trifluoromethylation of azomethine imines with trimethyl(trifluoromethyl)silane. *Angew. Chem. Int. Edn* **48,** 6324–6327 (2009).
82. Hintermann, L. & Togni, A. Catalytic enantioselective fluorination of β-ketoesters. *Angew. Chem. Int. Edn* **39,** 4359–4362 (2000).
83. Hamashima, Y., Yagi, K., Takano, H., Tamás, L. & Sodeoka, M. An efficient enantioselective fluorination of various α-ketoesters catalyzed by chiral palladium complexes. *J. Am. Chem. Soc.* **124,** 14530–14531 (2002).
84. Enders, D. & Hüttl, M. R. M. Direct organocatalytic α-fluorination of aldehydes and ketones. *Synlett* 991–993 (2005).
85. Marigo, M., Fielenbach, D., Braunton, A., Kjærsgaard, A. & Jørgensen, K. A. Enantioselective formation of stereogenic carbon–fluorine centers by a simple catalytic method. *Angew. Chem. Int. Edn* **44,** 3703–3706 (2005).
86. Steiner, D. D., Mase, N. & Barbas, C. F. III. Direct asymmetric α-fluorination of aldehydes. *Angew. Chem. Int. Edn* **44,** 3706–3710 (2005).
87. Beeson, T. D. & MacMillan, D. W. C. Enantioselective organocatalytic α-fluorination of aldehydes. *J. Am. Chem. Soc.* **127,** 8826–8828 (2005).
88. Ishimaru, T. *et al.* Cinchona alkaloid catalyzed enantioselective fluorination of allyl silanes, silyl enol ethers, and oxindoles. *Angew. Chem. Int. Edn* **47,** 4157–4161 (2008).
89. Kwiatkowski, P., Beeson, T. D., Conrad, J. C. & MacMillan, D. W. C. Enantioselective organocatalytic α-fluorination of cyclic ketones. *J. Am. Chem. Soc.* **133,** 1738–1741 (2011).
90. Hollingworth, C. *et al.* Palladium-catalyzed allylic fluorination. *Angew. Chem. Int. Edn* **50,** 2613–2617 (2011).
91. Katcher, M. H. & Doyle, A. G. Palladium-catalyzed asymmetric synthesis of allylic fluorides. *J. Am. Chem. Soc.* **132,** 17402–17404 (2010).
92. Eisenberger, P., Gischig, S. & Togni, A. Novel 10-I-3 hypervalent iodine-based compounds for electrophilic trifluoromethylation. *Chem. Eur. J.* **12,** 2579–2586 (2006).
93. Allen, A. E. & MacMillan, D. W. C. The productive merger of iodonium salts and organocatalysis: a non-photolytic approach to the enantioselective α-trifluoromethylation of aldehydes. *J. Am. Chem. Soc.* **132,** 4986–4987 (2010).
94. Beeson, T. D., Mastracchio, A., Hong, J., Ashton, K. & MacMillan, D. W. C. Enantioselective organocatalysis using SOMO activation. *Science* **316,** 582–585 (2007).
95. Nagib, D. A., Scott, M. E. & MacMillan, D. W. C. Enantioselective α-trifluoromethylation of aldehydes via photoredox organocatalysis. *J. Am. Chem. Soc.* **131,** 10875–10877 (2009).
Reports the first example of enantioselective, photoredox/organocatalytic α-trifluoromethylation of aldehydes.
96. Wu, X.-F., Anbarasan, P., Neumann, H. & Beller, M. From noble metal to Nobel Prize: palladium-catalyzed coupling reactions as key methods in organic synthesis. *Angew. Chem. Int. Edn* **49,** 9047–9050 (2010).
97. de Meijere, A. & Diederich, F. (eds) *Metal-Catalyzed Cross-Coupling Reactions* (Wiley, 2004).
98. Hartwig, J. F. Carbon–heteroatom bond-forming reductive elimination of amines, ethers, and sulfides. *Acc. Chem. Res.* **31,** 852–860 (1998).
99. Hartwig, J. F. Carbon–heteroatom bond formation catalysed by organometallic complexes. *Nature* **455,** 314–322 (2008).
100. Muci, A. R. & Buchwald, S. L. in *Topics in Current Chemistry* Vol. 219 (ed. Miyaura, N.) 131–209 (Springer, 2001).

Acknowledgements We thank the NSF (CHE-0952753) and the NIH-NIGMS (GM088237) for financial support.

Author Contributions T.R. developed the framework for the Review; all authors contributed sections.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to T.R. (ritter@chemistry.harvard.edu).

ARTICLE

doi:10.1038/nature10093

Royalactin induces queen differentiation in honeybees

Masaki Kamakura[1]

The honeybee (*Apis mellifera*) forms two female castes: the queen and the worker. This dimorphism depends not on genetic differences, but on ingestion of royal jelly, although the mechanism through which royal jelly regulates caste differentiation has long remained unknown. Here I show that a 57-kDa protein in royal jelly, previously designated as royalactin, induces the differentiation of honeybee larvae into queens. Royalactin increased body size and ovary development and shortened developmental time in honeybees. Surprisingly, it also showed similar effects in the fruitfly (*Drosophila melanogaster*). Mechanistic studies revealed that royalactin activated p70 S6 kinase, which was responsible for the increase of body size, increased the activity of mitogen-activated protein kinase, which was involved in the decreased developmental time, and increased the titre of juvenile hormone, an essential hormone for ovary development. Knockdown of epidermal growth factor receptor (Egfr) expression in the fat body of honeybees and fruitflies resulted in a defect of all phenotypes induced by royalactin, showing that Egfr mediates these actions. These findings indicate that a specific factor in royal jelly, royalactin, drives queen development through an Egfr-mediated signalling pathway.

Caste in social insects represents one of the major transitions from one level of organization to another in evolution[1]. The honeybee (*Apis mellifera*) exhibits polyphenism, that is, adult females form two interdependent castes, the queen and the worker, depending on their environment at critical periods of caste determination[2,3]. This dimorphism is not a consequence of genetic difference[4,5]. Queens have a larger body size and shorter developmental time than workers, have ten times the lifespan of workers, typically 1 to 2 years, and lay up 2,000 eggs per day, whereas workers rear young larvae and gather nectar[6,7]. When larvae are nourished with royal jelly, which is secreted by workers[2,3], they differentiate into queens. Royal jelly seems to contain a specific factor(s) that determines caste differentiation, but this has not previously been identified. Furthermore, the relationship between caste-specific modulation of juvenile hormone and ecdysteroid after ingestion of royal jelly and the developmental signal in caste differentiation has remained elusive. Therefore, I aimed to identify the factor(s) that induces caste differentiation in the honeybee and to investigate the mechanism through which this factor drives the caste-specific developmental pathway.

A caste differentiation-inducing factor in royal jelly

The dietary requirements for rearing queens are known[8], but a diet for rearing worker honeybees has not been reported. In connection with this, I found that larvae reared with royal jelly stored at 40 °C for 7 days, which did not exhibit any antifatigue effect[9], showed increased developmental times, decreased body weight at eclosion and decreased ovary size, compared to larvae fed a diet containing fresh royal jelly, even though they were queen–worker intermediates (Supplementary Fig. 1a–c). This result indicated that long-term storage of royal jelly at high temperature decreases the biological activity of royal jelly for queen differentiation. Therefore, royal jelly was stored at 40 °C for 7, 14, 21 and 30 days, and the effects of these royal jelly samples on caste differentiation were examined. Storage of royal jelly at 40 °C for up to 30 days caused a reduction in the growth of developing larvae, decreased weight at adult emergence, ovary size reduction and prolongation of the pre-adult development time in proportion to storage duration (Supplementary Fig. 1). Adult females reared with royal jelly stored at 40 °C for 30 days (40 °C/30 d royal jelly) developed with a full worker morphotype. These results indicate that the putative inducer of queen differentiation in royal jelly might be gradually degraded in proportion to the storage period at 40 °C, being completely degraded after 30 days. Therefore, the compositional changes in royal jelly during storage were investigated next.

First, the contents of several vitamins, 10-hydroxy-2-decenoic acid, carbohydrates and fatty acids in royal jelly samples stored at 4 °C and 40 °C for 30 days were measured. No significant differences were observed in the contents of the examined compounds, except pantothenic acid, which showed a decrease to 60% of the initial concentration during storage at 40 °C for 30 days (Supplementary Table 1). However, pantothenic acid did not induce the emergence of queens (data not shown), in agreement with a previous report[10]. Next, compositional changes of proteins in royal jelly during storage were analysed by means of high-performance liquid chromatography (HPLC) and native polyacrylamide gel electrophoresis (PAGE). A 450-kDa protein, a 170-kDa protein and a 57-kDa protein (designated as royalactin[11]) were degraded during storage (Supplementary Fig. 2a and Supplementary Fig. 3). Royalactin is a monomeric protein that exhibits epidermal growth factor (Egf)-like effects on rat hepatocytes[11,12]. The 170-kDa protein was completely degraded during storage at 40 °C for 14 days, being undetectable in royal jelly stored at 40 °C for 21 or 30 days; because this royal jelly can still influence ovary development and growth of developing larvae, the 170-kDa protein seems to be irrelevant to caste differentiation (Supplementary Figs 1 and 2a). Royalactin was degraded proportionally to the period of storage, and was completely lost during storage at 40 °C for 30 days, whereas only 10% of the 450-kDa protein was destroyed during storage at 40 °C for 30 days (Supplementary Fig. 2a).

Next, royalactin and the 450-kDa protein were purified (Supplementary Fig. 2b–d), and the effects of these factors on caste differentiation were examined in the same manner described above. As

[1]Biotechnology Research Center, Toyama Prefectural University, Imizu, Toyama 939-0398, Japan.

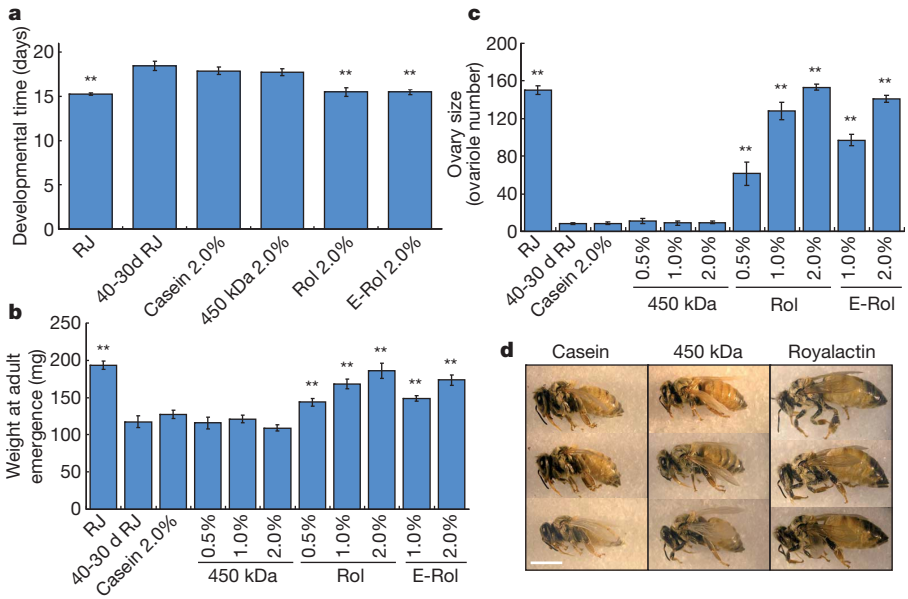


Figure 1 | Effects of casein, 450-kDa protein, royalactin and recombinant royalactin on caste characters in the honeybee. **a**–**c**, Developmental time (**a**), weight at adult emergence (**b**) and ovary size (**c**) in individuals ($n = 10$–28) reared with royal jelly (RJ), royal jelly stored at 40 °C for 30 days (40-30d RJ) or 40-30d RJ containing casein, 450-kDa protein, royalactin (RoI) or E-royalactin (E-RoI) were measured. **d**, Final adult size after eclosion is shown. Values are expressed as mean ± s.e.m. Values significantly different from those of larvae reared with 40-30d RJ are indicated by $^{**}P < 0.01$. Royalactin accounted for approximately 2.0% of RJ. Scale bar, 5 mm.

shown in Fig. 1, the 450-kDa protein (0.5% to 2.0% w/w diet) and casein (2.0% w/w diet), which was used as a control for evaluating nutritional effect, did not change the final adult size, developmental time, or ovary size in individuals reared with 40 °C/30 d royal jelly. In contrast, royalactin shortened developmental time and increased both weight at adult emergence and ovary size in proportion to the concentration added to a diet containing 40 °C/30 d royal jelly, and it induced larvae to develop into queens as effectively as did royal jelly at the concentration of 2.0% w/w diet (Fig. 1 and Supplementary Fig. 4a). Similar results were observed in larvae reared with recombinant royalactin (E-royalactin; 47 kDa), which was expressed in *Escherichia coli* and purified to homogeneity on SDS–PAGE (Fig. 1, Supplementary Fig. 2e and Supplementary Fig. 4a). Furthermore, royalactin and E-royalactin increased the juvenile hormone titre—which increases at the fourth larval instar to cause development into a queen[13,14]—in larvae given 40 °C/30 d royal jelly as potently as royal jelly, whereas the 450-kDa protein or casein had no effect (Supplementary Fig. 4b). Taken together, these results indicate that the stimulatory effect of royalactin on caste differentiation was not a nutritional effect but a morphogenic effect, and that royalactin is the major active factor in the induction of caste differentiation by royal jelly.

Effects of royal jelly and royalactin on *Drosophila*

Because no mutant stock of *Apis mellifera* has so far been developed, it is difficult to investigate the mechanism underlying honeybee caste differentiation at the individual level. On the other hand, fruitfly (*Drosophila melanogaster*), used as a model organism in many research fields, is available for genetic analysis in developmental biology. I considered that *Drosophila* might be suitable as a model insect for analysis of the mechanism of caste differentiation if royal jelly induced morphological and physiological changes in *Drosophila* similar to those induced in honeybee queens. Therefore, I investigated the influence of royal jelly on *Drosophila* larvae.

When *Drosophila* (Canton-S) larvae were reared with only royal jelly, they died before pupation (data not shown). However, *Drosophila* reared with medium containing 20% royal jelly, 8% yeast and 10% D-glucose had an increase in body size (body weight and body length) and fecundity, and had extended lifespan and shortened developmental time compared to flies reared with control medium or casein medium, which provide the same total energy as royal jelly medium (Fig. 2 and Supplementary Table 2). Furthermore, royal jelly medium increased cell size but not cell number (Supplementary Fig. 5). Royalactin increased body size, cell size and fecundity, extended lifespan and shortened developmental time in flies reared with 40 °C/30 d royal jelly (which did not influence morphological or physiological changes of flies), whereas 450-kDa protein or casein did not (Fig. 2, Supplementary Fig. 5 and Supplementary Table 2), in accordance with the observations that royalactin induced queen differentiation in honeybee as the major active factor in royal jelly. Thus, fresh royal jelly led genetically identical fly larvae to develop into adult individuals with phenotypes similar to queen bees, indicating that *Drosophila* could be used as a model insect for genetic analysis of caste differentiation.

Royalactin changes *Drosophila* phenotypes via Egfr

The insulin signalling pathway in metazoans has an important role in regulating body size, growth and metabolism[15,16]. First, I examined the effects of royal jelly on body size of *insulin receptor* (*InR*) mutants (InR^{E19}/InR^{E19} and InR^{p5545}/InR^{E19})[15] and mutant showing elevated levels of phosphatidylinositol-3 kinase (PI3K) activity in prothoracic gland and corpora allata with *P0206-Gal4* (*P0206>dP13K*)[16], all of which show reduced body size and weight. The *InR* mutants and *P0206>dPI3K* reared with royal jelly medium had larger body size and shorter developmental time than individuals reared with control medium or casein medium (Supplementary Fig. 6 and Supplementary Table 3).

I previously found that royalactin functions similarly to Egf in rat hepatocytes[11,12]. Therefore, I investigated the effects of royal jelly on body size of *Epidermal growth factor receptor* (*Egfr*) mutants ($Egfr^{tsla}/Egfr^{f24}$)[17]. Royal jelly did not influence body size or developmental time in the *Egfr* mutants (Supplementary Fig. 6 and Supplementary Table 3). Next, to determine the tissue specificity of royal jelly action in flies, I examined the influence of royal jelly on body size and developmental time in mutants in which expression of Egfr was silenced in the prothoracic gland, corpora allata or fat body, which are involved in body size regulation of *Drosophila*[16,18–21]. I used *Aug21-Gal4* (ref. 19) or *pumpless* (*ppl*)*-Gal4* (refs 16, 22) as a line with specific Gal4 expression in corpora allata or fat body, respectively. Royal jelly increased body size and shortened developmental time in *P0206>dEgfrRNAi*

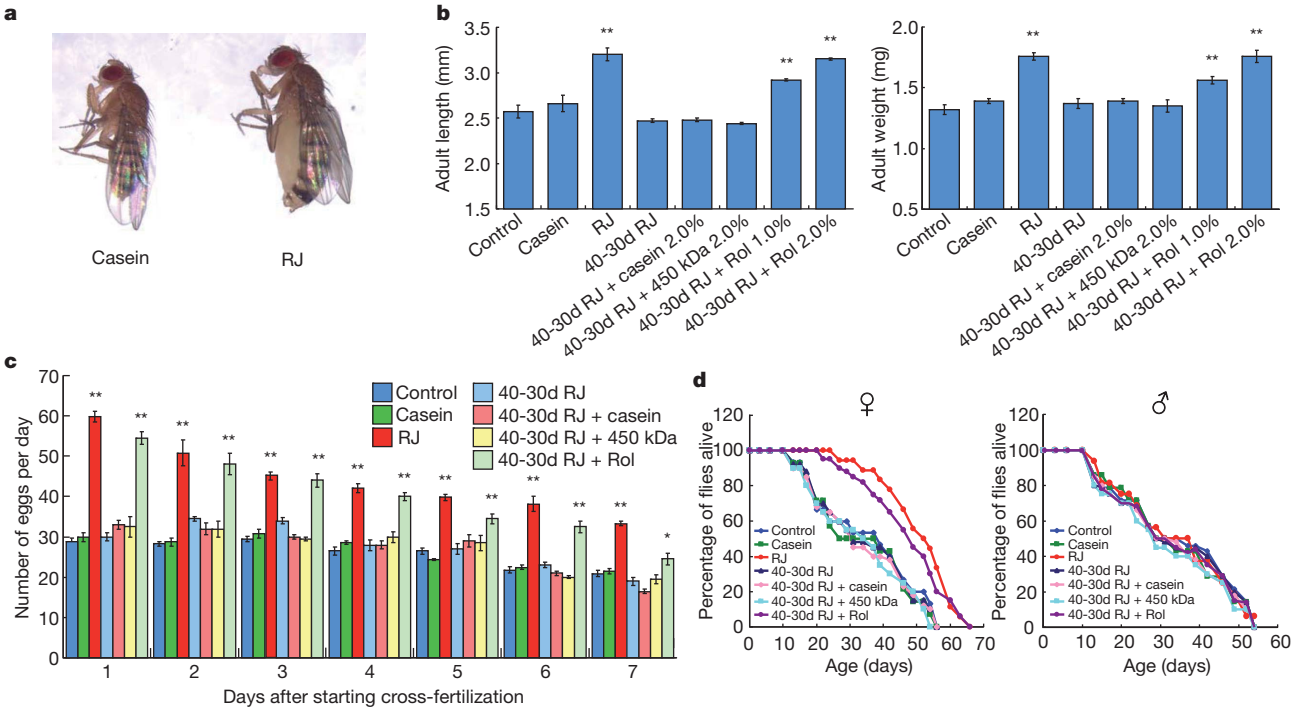


Figure 2 | Morphological and physiological changes of *Drosophila melanogaster* induced by royal jelly and royalactin. **a**, Body size of female adult flies reared with casein medium (8% yeast, 11.3% D-glucose, 2.8% casein, 1.3% D-fructose, 0.4% cornstarch, 0.76% soybean oil) and royal jelly medium (8% yeast, 10% D-glucose, 20% royal jelly). **b**–**d**, Body length (**b**, left), body weight (**b**, right), fecundity (**c**) and longevity (**d**) in wild-type (CS) fruit flies reared with control medium (8% yeast, 10% D-glucose), casein medium, royal jelly (RJ) medium or medium containing royal jelly stored at 40 °C for 30 days (40-30d RJ) or 40-30d RJ with casein, 450-kDa protein or royalactin (Rol) ($n > 40$). Values are expressed as mean ± s.e.m. Values significantly different from those of flies reared with control medium are indicated by $*P < 0.05$ or $**P < 0.01$.

flies (an RNA interference (RNAi) line for *Drosophila* Egfr) and *Aug21>dEgfrRNAi* flies, whereas it did not affect body size or developmental time in *ppl>dEgfrRNAi* flies (Supplementary Fig. 6 and Supplementary Table 3). Similar results were observed in *ppl>dEgfrRNAi* flies reared with royalactin (data not shown). I confirmed that *Drosophila* Egfr was expressed in the fat body of the wild-type flies (Supplementary Fig. 7). These findings demonstrated that not InR signalling, but rather Egfr signalling in the fat body was implicated in the increase of body size and reduction of developmental time by royal jelly or royalactin.

I next investigated how Egfr signals regulate changes of body size and developmental time in response to royal jelly. Royal jelly or royalactin activated S6K—which is activated by both phosphatidylinositol-dependent kinase 1 (PDK1) downstream of PI3K and target of rapamycin (TOR) downstream of PI3K/PDK1/Akt through stimulation of Egfr[23–26]—and mitogen-activated protein kinase (MAPK) in the larval fat body, and the activation of these enzymes by royalactin was suppressed by *Drosophila Egfr* RNAi in the fat body (Supplementary Fig. 8). Royal jelly did not increase the body size of *ppl>dPI3KDN* (*Drosophila* PI3K dominant-negative), *ppl>dPDK1RNAi*, *ppl>dAktRNAi*, *ppl>dTORDN* or *ppl>dS6KDN* flies, but shortened their developmental time, whereas *ppl>dRafRNAi* and *ppl>dMKP3* (ERK-inhibitory phosphatase)[27] reared with royal jelly showed increased body size but no early eclosion compared to the mutants reared with control medium or casein medium (Supplementary Fig. 6, Supplementary Table 3 and Supplementary Table 4). The increase of cell size in flies reared with royal jelly was repressed in *ppl>dEgfrRNAi* and *ppl>dS6KDN*, but not *ppl>dMKP3* (Supplementary Fig. 9). Loss of S6K function in *Drosophila* reduces body size by decreasing cell size but not cell number[28]. Activity of the MAPK pathway is reported to be unaffected by nutrients[29]. These results indicate that royalactin activated S6K through Egfr in the fat body, acting as a morphogenic factor to increase body size through an increase of cell size, and it also activated the MAPK pathway in the fat body to reduce the developmental time in *Drosophila*.

Drosophila phenotypes change in response to royalactin overexpression

To examine the stimulatory action of royalactin on *Drosophila* further, I examined the effect of overexpression of *royalactin* using the UAS/Gal4 system[30]. Surprisingly, *act>royalactin* showed increased body size, cell size, fecundity and longevity and shortened developmental time compared with *UAS-royalactin* (Fig. 3b–d, Supplementary Fig. 10 and Supplementary Table 5). Moreover, overexpression of *royalactin* specifically in the fat body or an Egfr signal using ppl-Gal4 or Gal4 driver of rhomboid (rho), which is the essential signal-generating component of Egfr signalling during development in *Drosophila*[31], induced the same phenotypes as *act>royalactin* (Fig. 3, Supplementary Fig. 10 and Supplementary Table 5). Royal jelly proteins were reported to contain royalactin, identical to major royal jelly protein (MRJP)1 and MRJP2–5 (ref. 32). I overexpressed *mrjp2–5* with *act-Gal4*, *rho-Gal4* and *ppl-Gal4*, and found that the body sizes of these mutants overexpressing *mrjp2–5* did not change (Supplementary Table 6). Overexpression of royalactin activated MAPK and S6K in the fat body of larvae, and this activation was inhibited by *Drosophila Egfr* RNAi (Supplementary Fig. 11). On the other hand, when *royalactin* was overexpressed with *P0206-Gal4* or *Aug21-Gal4*, it did not influence body size or developmental time of the mutants (Fig. 3b and Supplementary Table 5). Increase of body size and cell size in *ppl>royalactin* and *rho>royalactin* was suppressed by inhibition of Egfr and S6K, but not by abrogation of InR and MAPK (Fig. 3b, Supplementary Fig. 10 and data not shown). Reduction of developmental time in *ppl>royalactin* and *rho>royalactin* was repressed by inhibition of Egfr and MAPK, but not by inhibition of S6K (Supplementary Table 5 and data not shown). These results are consistent with the findings in flies reared with royal jelly.

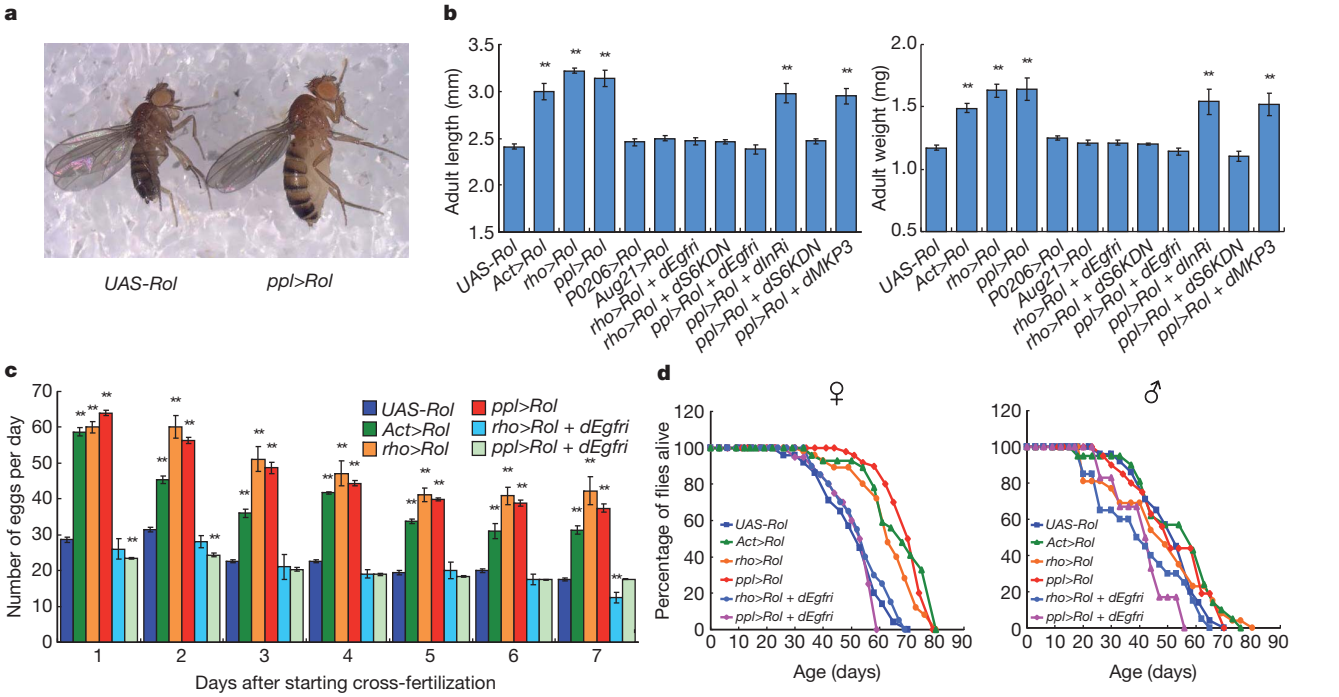


Figure 3 | Morphological and physiological changes of *Drosophila melanogaster* induced by overexpression of *royalactin*. **a**, Body size of female adult flies without or with overexpression of *royalactin* in the fat body (*UAS-Rol* or *ppl>Rol*). **b**, Body length (**b**, left) and body weight (**b**, right) in flies with overexpression of *royalactin* and in the signal factor suppression mutants in the *royalactin* overexpression background. $n > 40$. **c**, **d**, Fecundity (**c**) and longevity (**d**) in flies with overexpression of *royalactin* and in *Drosophila Egfr* interference (*dEgfri*) mutants in the *royalactin* overexpression background ($n > 50$). Values are expressed as mean ± s.e.m. Values significantly different from those of *UAS-royalactin* are indicated by $^{**}P < 0.01$.

Royalactin changes hormone metabolism in *Drosophila*

To investigate the relationship between the morphological and physiological changes induced by royalactin in flies and hormone modulation, I measured changes in the biosynthesis of a biologically active ecdysteroid, 20-hydroxyecdysone (20E), and juvenile hormone in wild-type flies given royal jelly during the larval period. Moreover, changes in gene expression of *yolk protein* (*yp*) during larval development were examined because juvenile hormone induces expression in the fat body of *yp*, which is essential for vitellogenesis, thereby promoting egg production in *Drosophila*[33]. Royal jelly and royalactin increased the 20E titre at 3 days after egg deposition (AED), and juvenile hormone titre and gene expression of *yp* at 4 days AED (Supplementary Fig. 12 and Supplementary Fig. 13). The increase of 20E titre in flies reared with royal jelly was suppressed in *ppl>dEgfrRNAi* and *ppl>dMKP3*, but not *ppl>dS6KDN* (Supplementary Fig. 14a), indicating that activation of MAPK downstream of Egfr in the fat body by royalactin induced 20E synthesis to shorten the developmental time. On the other hand, the increase of juvenile hormone titre, gene expression of *yp* and fecundity by royal jelly was repressed in *ppl>dEgfrRNAi*, but not in *ppl>dS6KDN* or *ppl>dMKP3* flies (Supplementary Fig. 14b–d and Supplementary Fig. 15). Because repression of MAPK in the fat body (*ppl>dMKP3*) did not abrogate the increase of *yp* expression and fecundity, the increase of 20E by royalactin seemed not to be associated with the increase of *yp* expression and oviposition. Taken together, these findings indicated that Egfr signalling in the fat body is activated by royalactin via a pathway distinct from that regulating body size and developmental time, leading to induction of juvenile hormone synthesis and a consequent increase of *yp* expression, thereby increasing fecundity. S6K in the fat body also seemed be associated only with the increase of body size by royal jelly.

On the other hand, increase of fecundity in flies with overexpression of *royalactin* was also repressed by *Drosophila Egfr* RNAi in the fat body but not by suppression of S6K and MAPK in the fat body (Fig. 3c and data not shown). These results were consistent with those obtained in flies reared with royal jelly. Increase of longevity induced by royal jelly was also abrogated in *ppl>dEgfrRNAi* flies, but not *ppl>dS6KDN* or *ppl>dMKP3* flies, indicating that Egfr in the fat body was essential for the increase of longevity in flies reared with royal jelly (Supplementary Fig. 14e and Supplementary Fig. 16a, b). Similar results were seen in the case of overexpression of *royalactin* (Fig. 3d and data not shown).

Suppression of queen differentiation in honeybees with RNAi

To confirm the signalling pathway involved in caste development, I reared honeybee larvae with suppression of *Apis mellifera InR* (*InR*) and *Egfr* by RNAi. Knockdown of *InR* did not affect final adult size, developmental time or ovary size in individuals reared with royal jelly, including a double-stranded RNA for green fluorescent protein (GFP), a control of RNAi, whereas *Egfr* RNAi reduced adult size and ovary size, and prolonged developmental time, compared with the control (GFP) (Fig. 4 and Supplementary Fig. 17a). These inhibitory effects of *Egfr* RNAi on queen differentiation were also observed in individuals reared with royalactin (data not shown). Royalactin activated MAPK and S6K through Egfr in fat body of honeybee larvae as effectively as did royal jelly (Supplementary Fig. 18). These results indicate that the activation of Egfr by royalactin is also involved in caste differentiation in the honeybee. Furthermore, suppression of honeybee *PI3K*, *PDK1*, *TOR* and *S6K* with RNAi inhibited the increase to final adult size induced by royal jelly, but did not affect changes of developmental time or ovary development (Fig. 4, Supplementary Fig. 17a and Supplementary Fig. 19). Royal jelly or royalactin increased the 20E titre in 3-day-old honeybee larvae, and the juvenile hormone titre and gene expression of *vitellogenin* (*vg*), a precursor of *yp*, in 4-day-old honeybee larvae given 40 °C/30 d royal jelly, whereas the 450-kDa protein and casein did not (Supplementary Fig. 4 and Supplementary Fig. 20). Increase of the 20E titre in honeybee larvae reared with royal jelly was abolished by *Egfr* RNAi and PD98059, a MAPK inhibitor, but not *S6K* RNAi (Supplementary Fig. 20a). PD98059 prolonged developmental time in larvae reared with royal jelly (data not shown). Increase

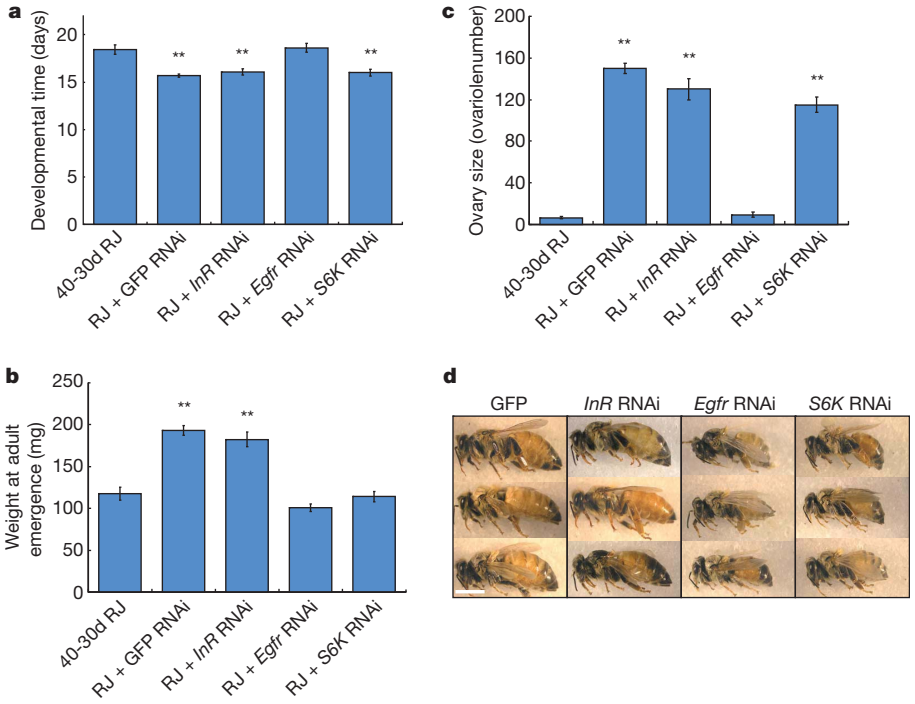


Figure 4 | Suppression of queen differentiation in honeybee with RNAi. **a**–**c**, Developmental time (**a**), weight at adult emergence (**b**) and ovary size (**c**) in individuals ($n = 10$–16) reared with royal jelly stored at 40 °C for 30 days (40-30d RJ) and royal jelly (RJ) containing dsRNA of GFP and signal factors. Values are expressed as mean ± s.e.m. Values significantly different from those of larvae reared with 40-30d RJ are indicated by $^{**}P < 0.01$. **d**, The final adult size after eclosion is shown in the photograph. Scale bar, 5 mm.

of juvenile hormone titre and *vg* expression in honeybee larvae reared with royal jelly was inhibited by *Egfr* RNAi but not *S6K* RNAi and PD98059 (Supplementary Fig. 20b, c). Thus, I found that activation of S6K by royalactin through Egfr was involved in the increase of body size in queens, whereas MAPK activity downstream of Egfr in response to royalactin was responsible for the increase of 20E synthesis, thereby shortening developmental time in the honeybee. Topical application of juvenile hormone to worker larvae results in the emergence of queen-like individuals with ovary development, but which display body sizes consistent with workers[34,35]. Therefore, an increase in juvenile hormone titre downstream of Egfr signalling activated by royalactin may have a function in ovary development in queens. These mechanisms are consistent with those of the morphological and physiological changes induced by royalactin in flies.

Here I provide the first evidence, to my knowledge, that royalactin acts on Egfr in the honeybee to induce queen differentiation. Furthermore, I found that administration and overexpression of royalactin in *Drosophila* caused morphological and physiological changes resembling the phenotypes of queen bee, through a similar mechanism to that of caste differentiation in the honeybee. These results provide new evidence that Egfr signalling has an important role in growth regulation.

The 450-kDa protein consists of apalbumin (420 kDa) and an oligomer of apisimin (5.5 kDa)[36]. Apalbumin is an oligomer of the gene product of *mrjp1* (ref. 36). On the other hand, royalactin is derived from *mrjp1* (ref. 37), but is present as a monomeric glycoprotein with a molecular mass of 57 kDa in royal jelly, and is structurally distinct from apalbumin (antibodies to royalactin do not recognize apalbumin)[11]. Apalbumin binds strongly to apisimin to form a stable complex (450-kDa protein); apalbumin was not separated from apisimin in the absence of detergent[36]. These results indicate that royalactin is not derived from apalbumin in royal jelly. The 40 °C/30 d royal jelly, which contained 90% of the initial concentration of 450-kDa protein, did not induce queen development, and the 450-kDa protein did not increase the rate of emergence of queens when it was added to a diet containing 40 °C/30 d royal jelly. However, both royalactin and E-royalactin induced queen differentiation in the honeybee. Thus, only royalactin, a monomer of MRJP1, functions as a caste determination factor. Royalactin induced prolonged longevity through Egfr in *Drosophila*, indicating that royalactin might have an important role in the prolongation of longevity in queens. To my knowledge, this is the first evidence that Egfr is involved in the regulation of longevity. Further research will be required to investigate the mechanism through which royalactin regulates lifespan in the fruitfly and the honeybee.

The association between royal jelly and caste formation has been known for more than 100 years, but the identity of the component(s) in royal jelly that induces queen development has been elusive. My results provide important insights into the process of caste development in the honeybee, and may also offer a valuable clue to eusociality and the evolution of social hymenopterans.

METHODS SUMMARY

Fly larvae were reared with medium containing royal jelly, D-glucose, yeast and agar at 25 °C. Honeybee larvae were reared with medium containing royal jelly, D-glucose, D-fructose and yeast extract at 34 °C with 96% humidity. Quantitative assay of juvenile hormone was carried out by high-resolution liquid chromatography-mass spectrometry (LC-MS) on a micrOTOF-Q instrument. The 20E titre of larvae was determined by the enzyme immunoassay (EIA) method. Quantitative analysis of gene expression was conducted by real-time PCR with the primers shown in Supplementary Tables 12 and 13. For honeybee RNAi experiments, the rearing diet containing enzymatically synthesized dsRNA at 150 μg ml^{-1} was administered to second instar larvae for 2 days.

Received 2 June 2010; accepted 5 April 2011.

Published online 24 April 2011; corrected 26 May 2011 (see full-text HTML version for details).

1. Maynard Smith, J. & Szathmary, L. *The Major Transitions in Evolution* (Freeman, 1995).
2. Haydak, M. H. Honey bee nutrition. *Annu. Rev. Entomol.* **15,** 143–156 (1970).
3. Patel, N. G., Haydak, M. H. & Gochnauer, T. A. Electrophoretic components of the proteins in honeybee larval food. *Nature* **186,** 633–634 (1960).
4. Weaver, N. Effects of larval age on dimorphic differentiation of the female honey bee. *Ann. Entomol. Soc. Am.* **50,** 283–294 (1957).
5. Shuel, R. W. & Dixon, S. E. The early establishment of dimorphism in the female honeybee, *Apis mellifera* L. *Insectes Soc.* **7,** 265–282 (1960).

6. Page, R. E. & Peng, C. Y. Aging and development in social insects with emphasis on the honey bee, *Apis mellifera* L. *Exp. Gerontol.* **36,** 695–711 (2001).
7. Wheeler, D. E. Developmental and physiological determinations of caste in social Hymenoptera: evolutionary implications. *Am. Nat.* **128,** 13–34 (1986).
8. Patel, A. *et al.* The making of a queen: TOR pathway is a key player in diphenic caste development. *PLoS ONE* **2,** e509 (2007).
9. Kamakura, M., Mitani, N., Fukuda, T. & Fukushima, M. Antifatigue effect of fresh royal jelly in mice. *J. Nutr. Sci. Vitaminol. (Tokyo)* **47,** 394–401 (2001).
10. Beetsma, J. The process of queen-worker differentiation in the honeybee. *Bee World* **60,** 24–39 (1979).
11. Kamakura, M., Suenobu, N. & Fukushima, M. 57-kDa protein in royal jelly enhances proliferation of primary cultured rat hepatocytes and increases albumin production in the absence of serum. *Biochem. Biophys. Res. Commun.* **282,** 865–874 (2001).
12. Kamakura, M. Signal transduction mechanism leading to enhanced proliferation of primary cultured adult rat hepatocytes treated with royal jelly 57-kDa protein. *J. Biochem.* **132,** 911–919 (2002).
13. Bloch, G., Wheeler, D. E. & Robinson, G. E. in *Hormones, Brain and Behavior* Vol. 3 (eds Pfaff, D. W., Arnold, A. P., Fahrbach, S. E., Etgen, A. M. & Rubin, R. T.) 195–235 (Academic Press, 2002).
14. Wirtz, P. & Beetsma, J. Induction of caste differentiation in the honeybee (*Apis mellifera* L.) by juvenile hormone. *Entomol. Exp. Appl.* **15,** 517–520 (1972).
15. Tatar, M. *et al.* A mutant *Drosophila* insulin receptor homolog that extends life-span and impairs neuroendocrine function. *Science* **292,** 107–110 (2001).
16. Colombani, J. *et al.* Antagonistic actions of ecdysone and insulins determine final size in *Drosophila*. *Science* **310,** 667–670 (2005).
17. Gilboa, L. & Lehmann, R. Soma-germline interactions coordinate homeostasis and growth in the *Drosophila* gonad. *Nature* **443,** 97–100 (2006).
18. Caldwell, P. E., Walkiewicz, M. & Stern, M. Ras activity in the *Drosophila* prothoracic gland regulates body size and developmental rate via ecdysone release. *Curr. Biol.* **15,** 1785–1795 (2005).
19. Mirth, C., Truman, J. W. & Riddiford, L. M. The role of the prothoracic gland in determining critical weight for metamorphosis in *Drosophila melanogaster*. *Curr. Biol.* **15,** 1796–1807 (2005).
20. Belgacem, Y. H. & Martin, J. R. Hmgcr in the *Corpus Allatum* controls sexual dimorphism of locomotor activity and body size via the insulin pathway in *Drosophila*. *PLoS ONE* **2,** e187 (2007).
21. Colombani, J. *et al.* A nutrient sensor mechanism controls *Drosophila* growth. *Cell* **114,** 739–749 (2003).
22. Zinke, I. *et al.* Suppression of food intake and growth by amino acids in *Drosophila*: the role of *pumpless*, a fat body expressed gene with homology to vertebrate glycine cleavage system. *Development* **126,** 5275–5284 (1999).
23. Navolanic, P. M., Steelman, L. S. & McCubrey, J. A. EGFR family signaling and its association with breast cancer development and resistance to chemotherapy. *Int. J. Oncol.* **22,** 237–252 (2003).
24. LeVea, C. M., Reeder, J. E. & Mooney, R. A. EGF-dependent cell cycle progression is controlled by density-dependent regulation of Akt activation. *Exp. Cell Res.* **297,** 272–284 (2004).
25. Rintelen, F., Stocker, H., Thomas, G. & Hafen, E. PDK1 regulates growth through Akt and S6K in *Drosophila*. *Proc. Natl Acad. Sci. USA* **98,** 15020–15025 (2001).
26. McManus, E. J. & Alessi, D. R. TSC1–TSC2: a complex tale of PKB-mediated S6K regulation. *Nature Cell Biol.* **4,** E214–E216 (2002).
27. Lee, K. S. *et al. Drosophila* short neuropeptide F signalling regulates growth by ERK-mediated insulin signalling. *Nature Cell Biol.* **10,** 468–475 (2008).
28. Montagne, J. *et al. Drosophila* S6 kinase: a regulator of cell size. *Science* **285,** 2126–2129 (1999).
29. Zhang, H. *et al.* Regulation of cellular growth by the *Drosophila* target of rapamycin *dTOR*. *Genes Dev.* **14,** 2712–2724 (2000).
30. Brand, A. H. & Perrimon, N. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development* **118,** 401–415 (1993).
31. Urban, S. Rhomboid proteins: conserved membrane proteases with divergent biological functions. *Genes Dev.* **20,** 3054–3068 (2006).
32. Schmitzova, J. *et al.* A family of royal jelly proteins of the honeybee *Apis mellifera* L. *Cell. Mol. Life Sci.* **54,** 1020–1030 (1998).
33. Bownes, M. The regulation of the yolk protein genes, a family of sex differentiation genes in *Drosophila melanogaster*. *Bioessays* **16,** 745–752 (1994).
34. Goewie, E. A. & Beetsma, J. Induction of caste differentiation in the honey bee (*Apis mellifera* L.) after topical application of JH-III. *Proc. K. Ned. Akad. Wet. C* **79,** 466–469 (1976).
35. Ebert, R. Influence of juvenile hormone on gravity orientation in the female honeybee larvae (*Apis mellifera* L.). *J. Comp. Physiol.* **137,** 7–16 (1980).
36. Bíliková, K. *et al.* Apisimin, a new serine-valine-rich peptide from honeybee (*Apis mellifera* L.) royal jelly: purification and molecular characterization. *FEBS Lett.* **528,** 125–129 (2002).
37. Kamakura, M. & Sakaki, T. A hypopharyngeal gland protein of the worker honeybee *Apis mellifera* L. enhances proliferation of primary-cultured rat hepatocytes and suppresses apoptosis in the absence of serum. *Protein Expr. Purif.* **45,** 307–314 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements I thank D. Yamamoto for provision of general fruitfly treatment methods and helpful advice; and S. Hayashi and T. Adachi-Yamada for instruction of dissection techniques in *Drosophila*. I thank T. Nonogaki and Y. Hasada for supply of honeybee larvae; K. Yu, M. Tatar, P. Leopold, G. Korge, Y. T. Ip, T. G. Wilson and D. Yamamoto for fly stocks. We are grateful to T. Oda for the gift of royal jelly, and to W. R. S. Steele for proofreading the article.

Author Contributions M.K. designed the research and performed the experiments. M.K. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.K. (kamakura@pu-toyama.ac.jp).

ARTICLE

doi:10.1038/nature10016

Probing cellular protein complexes using single-molecule pull-down

Ankur Jain[1], Ruijie Liu[2], Biswarathan Ramani[2], Edwin Arauz[3], Yuji Ishitsuka[4,5], Kaushik Ragunathan[1], Jeehae Park[1], Jie Chen[3], Yang K. Xiang[2] & Taekjip Ha[1,4,5]

Proteins perform most cellular functions in macromolecular complexes. The same protein often participates in different complexes to exhibit diverse functionality. Current ensemble approaches of identifying cellular protein interactions cannot reveal physiological permutations of these interactions. Here we describe a single-molecule pull-down (SiMPull) assay that combines the principles of a conventional pull-down assay with single-molecule fluorescence microscopy and enables direct visualization of individual cellular protein complexes. SiMPull can reveal how many proteins and of which kinds are present in the *in vivo* complex, as we show using protein kinase A. We then demonstrate a wide applicability to various signalling proteins found in the cytosol, membrane and cellular organelles, and to endogenous protein complexes from animal tissue extracts. The pulled-down proteins are functional and are used, without further processing, for single-molecule biochemical studies. SiMPull should provide a rapid, sensitive and robust platform for analysing protein assemblies in biological pathways.

Dynamic interactions between proteins guide almost every aspect of cellular function[1]. Understanding how the macromolecules in living cells interact holds the key to deciphering their roles in cellular function and regulation[2,3]. Individual proteins are also part of diverse sets of protein networks, making it challenging to tease apart various permutations of protein–protein interactions occurring in the cellular context[4]. Currently, the gold standard for determining interactions between proteins is the co-immunoprecipitation assay[5–7], which relies on affinity-based co-purification of interacting proteins, followed by identification via western blot or mass spectrometry. It is, however, difficult to determine how many copies of which proteins are present in the physiological complex using conventional immunoprecipitation. In addition, the many hours and multiple steps that often exist between sample preparation and measurements present uncertainties over the extent to which *in vivo* interactions are preserved before analysis.

In situ imaging methods based on resonance energy transfer[8,9], fluorescence correlation spectroscopy[10,11], two-hybrid methods[12,13] and the bimolecular fluorescence complementation assay[14] are other popular tools for studying pair-wise protein interactions. However, these methods cannot be applied to endogenous proteins and are, in general, blind to heterogeneous interactions between proteins and their stoichiometry.

Here we present a simple, direct and sensitive method to study cellular protein complexes with single-complex resolution. We call this method single-molecule pull-down or SiMPull because physiological macromolecular complexes are pulled down from cell or tissue extracts directly to the imaging surface of single-molecule fluorescence microscopy.

Experimental strategy and YFP pull-down

The key requirement for pull-down assays is the selective capture of a protein of interest (bait), which will bring along its binding partners (prey). We constructed a flow chamber using a microscope slide and a cover slip, passivated with methoxy polyethylene glycol (mPEG)[15] to prevent non-specific adsorption of cell extracts and antibodies, which should minimize false positives[7]. The imaging surface was also doped with biotinylated PEG and streptavidin, followed by biotinylated antibodies against the bait protein (Fig. 1a–d and Supplementary Fig. 1). When cell extracts are infused in the flow chamber, the surface-tethered antibody captures the bait protein together with any interacting partners. After washing away the unbound cell extract, co-immunoprecipitated prey proteins are visualized either through immunofluorescence labelling (Fig. 1a) or using genetically encoded fluorescent protein tags (Fig. 1b). This approach is extendable to multi-protein complexes via multi-colour labelling and has the potential to differentiate between multiple sub-complexes and configurations (Fig. 1c). When proteins are fluorescently labelled with a fixed ratio, photobleaching events yield stoichiometric information[16,17] (Fig. 1d).

We first validated the SiMPull assay for specific pull-down of yellow fluorescent protein (YFP) from cell extracts. When the crude lysate from cells overexpressing His_6-tagged YFP was infused into the flow chamber coated with anti-His antibody, we observed single YFP molecules (Fig. 1e, f), similar to the analysis performed using purified protein[18] (Supplementary Fig. 2). Binding of YFP to the antibody was stable over two hours (Supplementary Fig. 3). The blank slide surface showed ~30 fluorescent spots per imaging area, 2,500 μm^2, possibly due to surface impurities. The number of fluorescent spots per imaging area, N_f, due to specifically pulled-down proteins was 10–20 fold over the background; we could maintain a >10 fold signal-to-noise ratio by controlling the lysate dilution factor. Control experiments with YFP lysate on non-specific antibody and lysate without YFP expression showed the same N_f as blank. Even lysate with a tenfold higher concentration of YFP yielded only ~30 additional spots relative to the blank, implying a less than 0.5% contribution from non-specifically adsorbed proteins (Fig. 1e, f). N_f increased monotonically as the cell lysate concentration increased over three orders

[1]Center for Biophysics and Computational Biology and Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. [2]Department of Molecular and Integrative Physiology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. [3]Department of Cell and Developmental Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. [4]Department of Physics and Center for the Physics of Living Cells, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. [5]Howard Hughes Medical Institute, Urbana, Illinois 61801, USA.

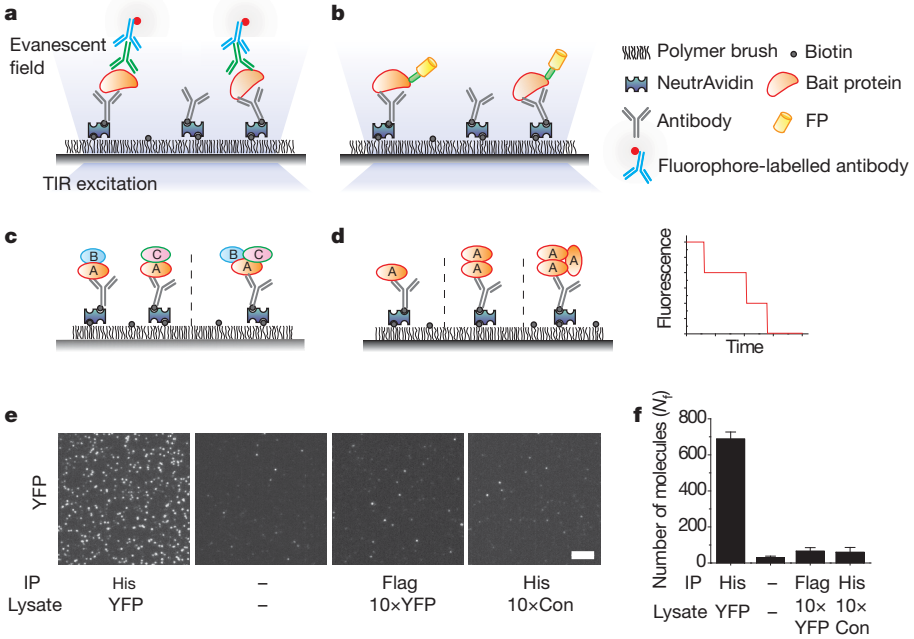


Figure 1 | Schematic for SiMPull assay. **a**, **b**, Immunoprecipitated protein complexes are visualized using TIRF microscopy using fluorophore-labelled antibody (**a**) or fluorescent protein (FP) tags (**b**). TIR, total internal reflection. **c**, Multi-colour colocalization can distinguish between subcomplexes (for example, AB + AC versus ABC). **d**, Photobleaching analysis can provide stoichiometric information. A simulated photobleaching trajectory for a trimeric protein. **e**, TIRF images for YFP pulled down from cells expressing His_6–YFP (YFP) and control cells (Con) using His-tag or a control (Flag-tag) antibody. Minus sign indicates no antibody or sample. IP, immunoprecipitate. Scale bar, 5 μm. **f**, Average number of fluorescent molecules per imaging area, N_f. Error bars denote standard deviation (s.d.) ($n > 20$).

of magnitude (Supplementary Fig. 4), showing that SiMPull can provide a quantitative estimate of protein concentration in cell lysate.

Single-molecule photobleaching analysis performed using the monomeric YFP and tandem dimeric YFP constructs (Supplementary Figs 5, 6) showed that accurate stoichiometric information can be obtained from the pulled-down proteins when we account for the fact that about 75% of YFP is fluorescently active[16].

Two-colour SiMPull of PKA complex

Next, we demonstrate the ability to pull-down single protein complexes from cell extracts using cyclic adenosine monophosphate (cAMP)-dependent protein kinase, protein kinase A (PKA), as our model system. PKA is a ubiquitous serine/threonine kinase that acts downstream of the G-protein coupled receptor (GPCR) pathway[19]. In the inactive state, PKA is a tetrameric complex consisting of two regulatory (R) and two catalytic (C) subunits. In the presence of cAMP, the complex dissociates, thereby activating the enzyme. We prepared C–HA–YFP and R–Flag–mCherry constructs (Fig. 2a).

When only C–HA–YFP was expressed in HEK293 cells, we could specifically pull-down the protein from the cell lysate using surface-immobilized antibodies against the HA or YFP epitope (Fig. 2b). As expected, these samples did not show any detectable fluorescence above background in the mCherry detection channel (Supplementary Fig. 7).

When the two subunits, R and C, were co-expressed, western blot showed that R and C co-immunoprecipitate[20] and they dissociate when cAMP is added to the lysate, confirming that the modified constructs retain the known properties of PKA (Fig. 2a). In a similar fashion, when we pulled down R–Flag–mCherry with anti-Flag antibody, we could detect both YFP and mCherry fluorescence spots. The number of fluorescent spots in mCherry and YFP channels was similar, indicating, on average, a one-to-one association (Fig. 2c, d). Fifty-seven per cent of YFP spots colocalized with a corresponding mCherry, as shown by individual images and their overlay (Fig. 2d). Incomplete colocalization may arise from basal tonic activation of PKA, inactive chromophores[16,21] or unbalanced expression of two proteins in individual cells. Adding cAMP analogue to the flow channel or pre-incubating the lysate with cAMP resulted in a greatly reduced number of C subunit (YFP spots) without a significant change in R subunit (mCherry spots) (Fig. 2c, e). After the reaction, only 4% of remaining YFP molecules colocalized with a corresponding mCherry.

Intracellular levels of cAMP can be increased by stimulating GPCRs for activation of adenylyl cyclases. When the cells overexpressing the PKA complex were stimulated with forskolin, an agonist of

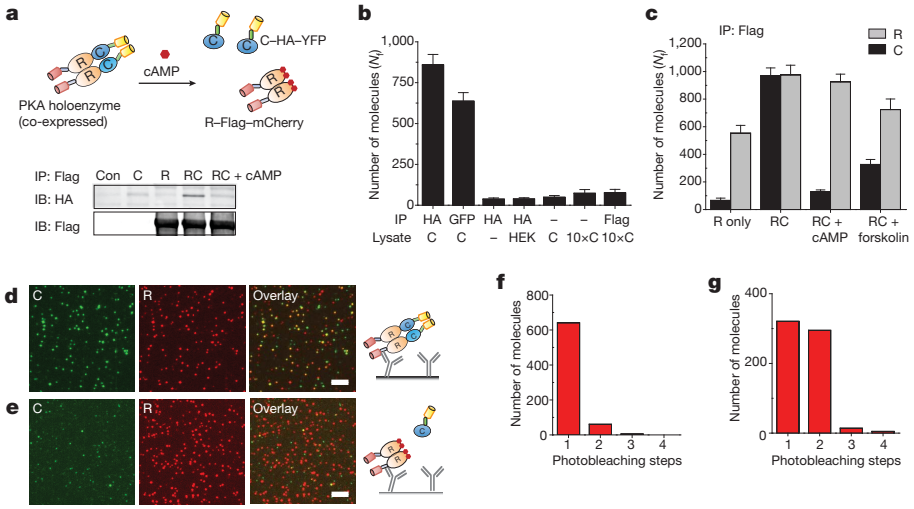


Figure 2 | PKA pull-down. **a**, Schematic of PKA construct. In western blot, C–HA–YFP is pulled down via R–Flag–mCherry; on adding cAMP, PKA dissociates. IB, immunoblot. **b**, N_f for C–HA–YFP (C) as a function of lysates and antibodies demonstrate the specificity of pull-down. **c**–**e**, PKA complex pull-down. **c**, N_f for YFP (C) and mCherry (R) spots. **d**, Images of single PKA complexes, YFP (left), mCherry (centre) and overlay (right). **e**, On adding cAMP, YFP spots decrease significantly. Scale bars, 5 μm. **f**, **g**, Photobleaching step distribution (**f**) for C–HA–YFP-only lysate and (**g**) for C–HA–YFP pulled down via R–Flag–mCherry. Error bars denote s.d. ($n > 20$).

adenylyl cyclase for cAMP production, the amount of active PKA in cells increased, as evidenced by dissociation of PKA in a western blot (Supplementary Fig. 8). Similarly, in our SiMPull assay, the number of C subunits (YFP molecules) pulled down through R subunits decreased significantly (Fig. 2c).

We explored the stoichiometry of PKA using photobleaching analysis. When C–HA–YFP was expressed alone and pulled down using HA antibody, 91% of YFP traces exhibited one-step photobleaching, indicating a monomeric population (Fig. 2f). When C–HA–YFP and R–Flag–mCherry were co-expressed and pulled down using anti-Flag antibody, 47% of YFP traces showed two photobleaching steps (Fig. 2g) and 51% of molecules bleached in one step. Assuming a 75% active fraction of YFP[16], this is consistent with the known stoichiometry of two catalytic subunits in each PKA. We did not perform photobleaching-based stoichiometry analysis for mCherry owing to its inferior photophysical properties[22].

Applications of SiMPull

Next, we examined the applicability of SiMPull to protein complexes from various cellular environments using different capture and detection configurations (Fig. 3).

Receptor pull-down

Membrane protein complexes are particularly difficult to analyse using conventional methods, thus motivating new approaches[23,24]. Their stoichiometry cannot be determined using photobleaching in the cell unless the areal density of the protein complex is low enough for single-molecule detection[16,25,26]. SiMPull should be able to detect individual complexes if membrane patches containing one complex can be isolated. As a test, we applied SiMPull to the β_2-adrenergic receptor (β_2AR), a prototypical GPCR. HEK293 cells were transfected with Flag–YFP–β_2AR and membrane proteins were solubilized. We could specifically pull-down the receptor using antibodies against YFP or Flag (Fig. 3a–c). Twenty-nine per cent of the traces showed two distinct bleaching steps (Supplementary Fig. 9a), indicating a ~51% dimer population, assuming 75% of active fluorophores. Less than 3% of the traces showed three or more photobleaching steps. Our observation of β_2AR homodimerization is consistent with previous studies[27,28]. To test if β_1ARs might form hetero-oligomers with β_2ARs[28], we co-expressed mCherry–β_1AR and YFP–β_2AR. Using antibodies against mCherry–β_1AR, we could pull-down YFP–β_2AR and vice versa (Supplementary Fig. 9b–e). The two fluorophores colocalized with ~42% overlap, consistent with hetero-oligomer formation.

Mitochondrial protein pull-down

Mitochondrial antiviral signalling (MAVS) is a mitochondrial outer membrane protein involved in the innate immune response[29]. When isolated mitochondrial fractions from cells overexpressing YFP–MAVS[30] were applied to a surface with anti-YFP, we observed bright fluorescent structures (Fig. 3e, left), indicating the presence of several YFP molecules. On pre-solubilizing the mitochondrial preparation using mild detergent, we observed isolated single YFP spots (Fig. 3e, centre), 86% of which showed single photobleaching steps (Supplementary Fig. 10), supporting the monomeric state of solubilized MAVS. This observation indicates that the bright fluorescent structures observed in unsolubilized preparations are due to immunoprecipitated mitochondrial membrane patches or whole mitochondria. It may be possible to specifically immobilize cellular organelles or their components using antibodies against suitable marker proteins and perform single-molecule measurements in a physiologically relevant context.

Immunofluorescence detection of single complexes

We extended the assay to detection via antibodies using mammalian target of rapamycin complex 1 (mTORC1) as a model system. mTORC1 is a key signalling complex that regulates cell growth and

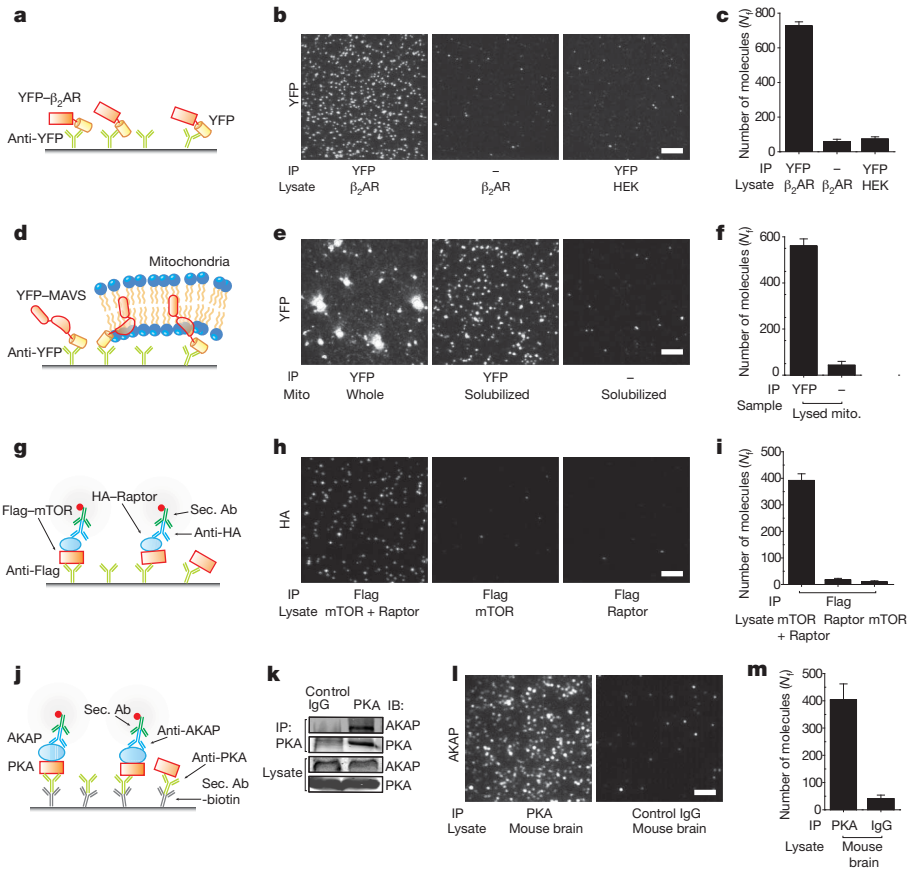


Figure 3 | Applications of SiMPull assay. **a–c**, β_2AR–YFP pull-down. **d–f**, MAVS pull-down. Mitochondrial fraction (mito.) from cells overexpressing YFP–MAVS was added either directly or after detergent solubilization. **g–i**, mTORC1 pull-down. Lysate from cells expressing Flag–mTOR, HA–Raptor or both was applied on chambers with Flag antibody, and probed through primary antibody against HA and labelled secondary antibody (Sec. Ab). **j–m**, Endogenous PKA–AKAP complex pull-down from mouse brain extract. **k**, Western blot shows AKAP immunoprecipitation with PKA antibody. **l**, Immunofluorescence images of AKAP150 pulled down through PKA antibody. **c**, **f**, **i**, **m** show N_f. Scale bars, 5 μm. Error bars denote s.d. ($n > 20$).

metabolism in response to nutrient availability[31,32]. In addition to mammalian target of rapamycin (mTOR), a defining component of mTORC1 is regulatory associated protein of mTOR (Raptor; also known as RPTOR), which associates with mTOR at an equimolar ratio[31,33]. We expressed Flag–mTOR and HA–Raptor in HEK293 cells. Flag–mTOR was pulled down using biotinylated Flag antibody; Raptor was detected using HA antibody followed by fluorescently labelled secondary antibody (Fig. 3g). When both Flag–mTOR and HA–Raptor were co-expressed, we observed the detection antibody binding as fluorescent spots whereas background level of fluorescence was detected when only one of the two proteins was expressed (Fig. 3h, i), demonstrating antibody-based detection in SiMPull.

Pull-down of endogenous complexes in native tissues

Exogenous expression may lead to non-physiological associations between proteins. Pull-down of endogenously expressed proteins, although desirable, is challenging owing to low abundance, high background interaction with other cellular proteins, and general lack of high-affinity antibodies. We tested if SiMPull can be used to detect interactions between endogenous proteins. A kinase anchoring proteins (AKAPs) bind to PKA and confine it to discrete locations in the cell[34]. Figure 3k shows AKAP150 can be co-immunoprecipitated with PKA from mouse brain extract.

Primary antibodies against proteins are often expensive and difficult to label with biotin or fluorophores. Thus, to keep our approach general, we used biotin-labelled secondary antibody to immobilize the antibody against the bait (PKA), and applied mouse brain extract. On probing for the prey protein (AKAP150) using its primary antibody and fluorescently labelled secondary antibody, we observed tenfold more fluorescent spots in the channel with PKA antibody as compared to the control channel (Fig. 3l, m). SiMPull required a 20-fold lower sample volume as compared to the corresponding western blot. This sensitivity allowed detection of PKA–AKAP binding from mouse heart tissue, which was below the detection limit of the conventional western blot under the same conditions (Supplementary Fig. 11).

SiMPull as a preparatory tool

A key advantage of SiMPull is that protein complexes can be directly observed from a fresh cell lysate, bypassing purification procedures. We tested if SiMPull can be used for functional analysis of pulled-down proteins. PcrA, a superfamily 1 helicase, is an ATP-driven motor protein that binds and translocates on single-stranded DNA (ssDNA)[35]. His_6-tagged PcrA was pulled down from bacterial lysate using anti-His antibody and fluorescently labelled DNA molecules were added to the flow channel (Fig. 4a). Fluorescent spots due to labelled DNA binding appeared in the flow channel with pulled-down PcrA, whereas the control channel showed minimal DNA binding (Fig. 4b, c).

When PcrA binds to a partial duplex DNA with a 5′ overhang, it anchors itself to the junction and repetitively reels in the ssDNA[35]. By labelling the DNA with a donor at the tail end and an acceptor at the junction (Fig. 4a), we could observe the reeling-in activity as a gradual increase in fluorescence resonance energy transfer (FRET) (Fig. 4d). Once PcrA reaches the end of the ssDNA, it runs off the ssDNA track and repeats the process from the junction over and over, resulting in cyclic increases and decreases in FRET. Eighty-six per cent (161 of 188) of bound FRET-labelled DNA molecules exhibited repetitive cycling. On increasing the ATP concentration, translocation became faster (Supplementary Fig. 12a), and in the absence of ATP, DNA remained bound but no reeling-in activity was observed (Supplementary Fig. 12b). The mean translocation time matched well with the data obtained with purified protein (Fig. 4e, f). Thus, SiMPull can pull-down functional macromolecules directly from cell extracts for subsequent single-molecule biochemistry *in situ*.

Discussion

We have established a single-molecule platform for analysing the cellular association of macromolecules. SiMPull can be used as an extension of commonly used western blot analysis without requiring additional sample preparation (Supplementary Fig. 13) and confers several key advantages. First, it can provide quantitative data on sub-populations of different association states. Second, it provides information on complex stoichiometry if the proteins can be stoichiometrically labelled. Third, the high sensitivity allows the study of complexes of low abundance, and a suitable calibration (Supplementary Fig. 4) should make it possible to determine the expression level of protein complexes in cell lysate. Fourth, the whole assay took about 30 min, considerably shorter than conventional western blot. In a pilot experiment, we could dilute the cell lysate, pull-down and quantify YFP in 2.5 min (Supplementary Fig. 14). Therefore, it should be possible to analyse even relatively weak protein complexes as long as the dissociation rate constant k_{off} is equal to or smaller than 0.01 s^{-1}. Combining this with microfluidics platform, cross-linking methods or zero mode waveguide[36] may extend the method to complexes with even higher k_{off}.

Cellular processes are under tight spatiotemporal regulation. To overcome ensemble averaging over heterogeneous cell populations, SiMPull may be combined with fluorescence-aided cell sorting to selectively analyse a subpopulation or may even be pushed to the single-cell level. In a preliminary experiment, we could pull-down and quantify proteins from 10 cells obtained through cell sorting (Supplementary Fig. 15) compared to the ~5,000 cells usually required for a western blot[37]. Single-cell SiMPull may enhance the recently demonstrated ability to quantify proteins and RNA numbers in single cells[38]. As in conventional western blot, the sensitivity and specificity of SiMPull are determined by the quality of capture and detection antibodies. The assay may be combined with recent developments in labelling

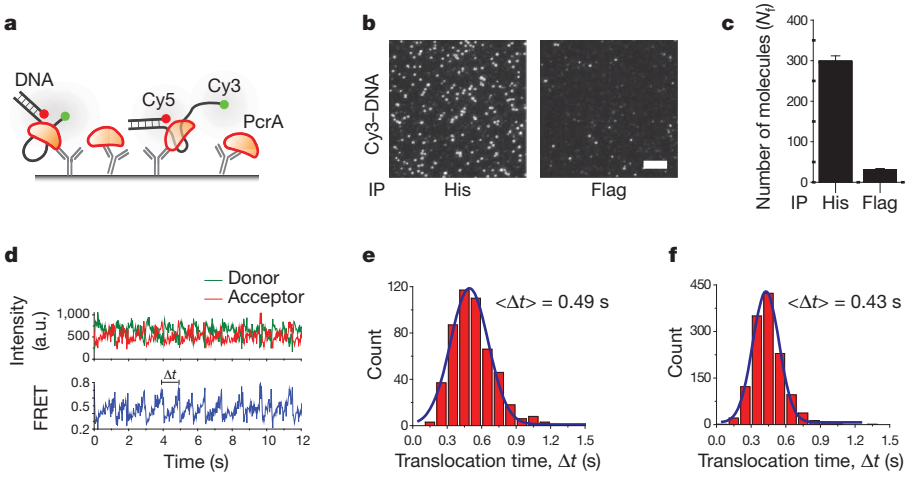


Figure 4 | PcrA pull-down and activity. **a**, Schematic. **b**, **c**, Labelled DNA binding to immunoprecipitated PcrA. Scale bar, 5 μm. Error bars represent s.d. ($n > 20$). **d**, A typical time trace of repetitive reeling-in activity of PcrA monitored by FRET. a.u., arbitrary units. **e**, **f**, The distribution of translocation times, Δt, and its mean, $<\Delta t>$, for purified PcrA (**e**) and for PcrA pulled down from cell extracts (**f**), at 1 mM ATP concentration.

strategies[39] for further improvement in sensitivity and labelling efficiency.

Post-translational modifications have an important role in cellular processes but are difficult to reproduce in recombinant proteins. In addition, the necessary co-factors or ligands for a protein of interest are often unknown. SiMPull as a preparatory tool provides a possibility to study these modified proteins or protein complexes that cannot be purified using conventional methods.

METHODS SUMMARY

Flow chambers were prepared on mPEG passivated quartz slides doped with biotin PEG[15]. Biotinylated antibodies were immobilized by incubating ~10 nM of antibody for 10 min on NeutrAvidin (Thermo) coated flow chambers. A prism type total internal reflection fluorescence (TIRF) microscope was used to acquire the single-molecule data[40]. Samples with fluorescent protein tag were serially diluted to obtain well-isolated spots on the surface upon 20 min of incubation over immobilized antibody surface. All dilutions were made immediately before addition to the flow chamber in 10 mM Tris-HCl pH 8.0, 50 mM NaCl buffer with 0.1 mg ml^{-1} bovine serum albumin (New England Biolabs), unless specified. Unbound antibodies and sample were removed from the channel by washing with buffer twice between successive additions. For immunofluorescence detection, immunoprecipitated complexes were incubated with a different antibody against prey protein (~10 nM) for 20 min and fluorescent-dye-labelled secondary antibody (2–5 nM) for 5 min before imaging. Single-molecule analysis was performed using scripts written in Matlab.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 2 December 2010; accepted 21 March 2011.

1. Alberts, B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92,** 291–294 (1998).
2. Papin, J. A., Hunter, T., Palsson, B. O. & Subramaniam, S. Reconstruction of cellular signalling networks and analysis of their properties. *Nature Rev. Mol. Cell Biol.* **6,** 99–111 (2005).
3. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415,** 141–147 (2002).
4. Yamada, T. & Bork, P. Evolution of biomolecular networks: lessons from metabolic and protein interactions. *Nature Rev. Mol. Cell Biol.* **10,** 791–803 (2009).
5. Barrios-Rodiles, M. *et al.* High-throughput mapping of a dynamic signaling network in mammalian cells. *Science* **307,** 1621–1625 (2005).
6. Puig, O. *et al.* The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* **24,** 218–229 (2001).
7. Gingras, A. C., Gstaiger, M., Raught, B. & Aebersold, R. Analysis of protein complexes using mass spectrometry. *Nature Rev. Mol. Cell Biol.* **8,** 645–654 (2007).
8. Wallrabe, H. & Periasamy, A. Imaging protein molecules using FRET and FLIM microscopy. *Curr. Opin. Biotechnol.* **16,** 19–27 (2005).
9. Carriba, P. *et al.* Detection of heteromerization of more than two proteins by sequential BRET-FRET. *Nature Methods* **5,** 727–733 (2008).
10. Slaughter, B. D., Schwartz, J. W. & Li, R. Mapping dynamic protein interactions in MAP kinase signaling using live-cell fluorescence fluctuation spectroscopy and imaging. *Proc. Natl Acad. Sci. USA* **104,** 20320–20325 (2007).
11. Zamir, E., Lommerse, P. H., Kinkhabwala, A., Grecco, H. E. & Bastiaens, P. I. Fluorescence fluctuations of quantum-dot sensors capture intracellular protein interaction dynamics. *Nature Methods* **7,** 295–298 (2010).
12. Fields, S. & Song, O. A novel genetic system to detect protein–protein interactions. *Nature* **340,** 245–246 (1989).
13. Eyckerman, S. *et al.* Design and application of a cytokine-receptor-based interaction trap. *Nature Cell Biol.* **3,** 1114–1119 (2001).
14. Kerppola, T. K. Bimolecular fluorescence complementation (BiFC) analysis as a probe of protein interactions in living cells. *Annu. Rev. Biophys.* **37,** 465–487 (2008).
15. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nature Methods* **5,** 507–516 (2008).
16. Ulbrich, M. H. & Isacoff, E. Y. Subunit counting in membrane-bound proteins. *Nature Methods* **4,** 319–321 (2007).
17. Reyes-Lamothe, R., Sherratt, J. D. & Leake, M. C. Stoichiometry and architecture of active DNA replication machinery in *Escherichia coli*. *Science* **328,** 498–501 (2010).
18. Mashanov, G. I., Tacon, D., Knight, A. E., Peckham, M. & Molloy, J. E. Visualizing single molecules inside living cells using total internal reflection fluorescence microscopy. *Methods* **29,** 142–152 (2003).
19. Collins, S., Caron, M. G. & Lefkowitz, R. J. Regulation of adrenergic receptor responsiveness through modulation of receptor gene expression. *Annu. Rev. Physiol.* **53,** 497–508 (1991).
20. Taylor, S. S. *et al.* PKA: a portrait of protein kinase dynamics. *Biochim. Biophys. Acta* **1697,** 259–269 (2004).
21. Maeder, C. I. *et al.* Spatial regulation of Fus3 MAP kinase activity through a reaction-diffusion mechanism in yeast pheromone signalling. *Nature Cell Biol.* **9,** 1319–1326 (2007).
22. Yu, Y. *et al.* Structural and molecular basis of the assembly of the TRPP2/PKD1 complex. *Proc. Natl Acad. Sci. USA* **106,** 11558–11563 (2009).
23. Lopez-Gimenez, J. F., Canals, M., Pediani, J. D. & Milligan, G. The α_{1b}-adrenoceptor exists as a higher-order oligomer: effective oligomerization is required for receptor maturation, surface delivery, and function. *Mol. Pharmacol.* **71,** 1015–1029 (2007).
24. Schwarzenbacher, M. *et al.* Micropatterning for quantitative analysis of protein–protein interactions in living cells. *Nature Methods* **5,** 1053–1060 (2008).
25. Yu, J., Xiao, J., Ren, X., Lao, K. & Xie, X. S. Probing gene expression in live cells, one protein molecule at a time. *Science* **311,** 1600–1603 (2006).
26. Leake, M. C. *et al.* Stoichiometry and turnover in single, functioning membrane protein complexes. *Nature* **443,** 355–358 (2006).
27. Angers, S. *et al.* Detection of β_2-adrenergic receptor dimerization in living cells using bioluminescence resonance energy transfer (BRET). *Proc. Natl Acad. Sci. USA* **97,** 3684–3689 (2000).
28. Mercier, J. F., Salahpour, A., Angers, S., Breit, A. & Bouvier, M. Quantitative assessment of β_1- and β_2-adrenergic receptor homo- and heterodimerization by bioluminescence resonance energy transfer. *J. Biol. Chem.* **277,** 44925–44931 (2002).
29. Seth, R. B., Sun, L., Ea, C. K. & Chen, Z. J. Identification and characterization of MAVS, a mitochondrial antiviral signaling protein that activates NF-κB and IRF3. *Cell* **122,** 669–682 (2005).
30. Baril, M., Racine, M. E., Penin, F. & Lamarre, D. MAVS dimer is a crucial signaling component of innate immunity and the target of hepatitis C virus NS3/4A protease. *J. Virol.* **83,** 1299–1311 (2009).
31. Kim, D. H. *et al.* mTOR interacts with raptor to form a nutrient-sensitive complex that signals to the cell growth machinery. *Cell* **110,** 163–175 (2002).
32. Sabatini, D. M. mTOR and cancer: insights into a complex relationship. *Nature Rev. Cancer* **6,** 729–734 (2006).
33. Yip, C. K., Murata, K., Walz, T., Sabatini, D. M. & Kang, S. A. Structure of the human mTOR complex I and its implications for rapamycin inhibition. *Mol. Cell* **38,** 768–774 (2010).
34. Tunquist, B. J. *et al.* Loss of AKAP150 perturbs distinct neuronal processes in mice. *Proc. Natl Acad. Sci. USA* **105,** 12557–12562 (2008).
35. Park, J. *et al.* PcrA helicase dismantles RecA filaments by reeling in DNA in uniform steps. *Cell* **142,** 544–555 (2010).
36. Levene, M. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299,** 682–686 (2003).
37. Schulte, R., Talamas, J., Doucet, C. & Hetzer, M. W. Single bead affinity detection (SINBAD) for the analysis of protein–protein interactions. *PLoS ONE* **3,** e2061 (2008).
38. Taniguchi, Y. *et al.* Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329,** 533–538 (2010).
39. Chen, I. & Ting, A. Y. Site-specific labeling of proteins with small molecules in live cells. *Curr. Opin. Biotechnol.* **16,** 35–40 (2005).
40. Myong, S., Rasnik, I., Joo, C., Lohman, T. M. & Ha, T. Repetitive shuttling of a motor protein on DNA. *Nature* **437,** 1321–1325 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Myong, P. Jena, S. Arslan and R. Vafabakhsh for discussions. The expression vector encoding the *YFP-MAVS* gene was a gift from D. Lamarre. This work was funded by NIH grants (AI083025, GM065367 to T.H.; HL082846 to Y.K.X.; AR048914 to J.C.). Additional support was provided by NSF grants (0646550, 0822613 to T.H.). T.H. is an investigator with the Howard Hughes Medical Institute.

Author Contributions A.J., Y.K.X. and T.H. designed the research. A.J., R.L. and Y.I. conducted experiments, R.L., B.R., E.A., J.C. and J.P. provided samples, K.R. and Y.I. contributed important ideas to the experiments, A.J. and R.L. analysed the data and A.J., Y.K.X. and T.H. wrote the paper with inputs from other authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to T.H. (tjha@illinois.edu) or Y.K.X. (kevinyx@illinois.edu).

METHODS

Overview. Flow chambers were prepared on mPEG passivated quartz slides doped with biotin PEG[15]. Biotinylated antibodies were immobilized by incubating ~10 nM of antibody for 10 min on NeutrAvidin (Thermo) coated flow chambers. A prism type total internal reflection fluorescence (TIRF) microscope was used to acquire the single-molecule data[40]. Samples with fluorescent protein tag were serially diluted to obtain well-isolated spots on the surface upon 20 min of incubation over immobilized antibody surface. All dilutions were made immediately before addition to the flow chamber in 10 mM Tris-HCl pH 8.0, 50 mM NaCl buffer with 0.1 mg ml^{-1} bovine serum albumin (New England Biolabs), unless specified. Unbound antibodies and sample were removed from the channel by washing with buffer twice between successive additions. For immunofluorescence detection, immunoprecipitated complexes were incubated with a different antibody against prey protein (~10 nM) for 20 min and fluorescent-dye-labelled secondary antibody (2–5 nM) for 5 min before imaging. Single-molecule analysis was performed using scripts written in Matlab.

Single-molecule imaging and spot counting. A prism type TIRF microscope was used to acquire single-molecule data[40]. YFP was excited at 488 nm; mCherry was excited at 532 or 568 nm. Narrow band-pass filters were used to avoid cross-talk between channels (HQ 535/30 from Chroma Technology for YFP and BL 607/36 from Semrock for mCherry). All experiments were performed at room temperature (22–25 °C) unless specified. Single-molecule analysis was performed as described earlier[15]. Mean spot count per image (imaging area 2,500 μm^2) and standard deviation were calculated from images taken from 20 or more different regions.

Photobleaching analysis. Single-molecule fluorescence time traces of surface immobilized YFP-tagged proteins were manually scored for the number of bleaching steps[16]. To avoid false colocalization, samples were immobilized at an optimal surface density (~300 molecules in 2,500 μm^2 imaging area). The number of photobleaching steps (single frame intensity drops of equal size) in each trace was manually determined, following published procedures[16]. The fluorescence trace of each molecule was classified as having 1–4 bleaching steps or was discarded if no clean bleaching steps could be identified (Supplementary Fig. 5). Some fluorescent protein molecules exhibited blinking, but under most circumstances distinct fluorescence intensity levels could be readily determined despite this blinking behaviour (Supplementary Fig. 5a, b). Separate counts were maintained for each case. At least 500 molecules were analysed for each sample. The probability of missed bleaching events due to simultaneous bleaching of both fluorescent proteins within the same imaging window is ~5%. The population distribution of observed bleaching events and discarded traces is reported in Supplementary Table 1. For future extensions to complexes with many more copies of the same protein, automated algorithms for scoring photobleaching steps would be required[41].

Single-molecule colocalization. Colocalization between YFP and mCherry was performed using a method similar to that described previously[42]. Briefly, we took two separate movies of the same region using YFP and mCherry excitation. The fluorescent spots in both images were fit with Gaussian profiles to determine the centre positions of the molecules to half-pixel accuracy. Next, for each molecule in the YFP image, we determined the mCherry molecules with their centre within a 2-pixel (~300 nm) distance. The number of molecules where this colocalization occurred divided by the total number of YFP molecules was presented as overlap percentage.

YFP constructs and pull-down. As YFP has been shown to dimerize, monomeric YFP was generated through site-directed mutagenesis of alanine 207 to lysine using pEYFP-C1 as the DNA template (Clontech). For bacterial expression, monomeric YFP was cloned into the SalI and XhoI sites of the pET-28b$^+$ vector. BL21 DE3 cells were transformed with the YFP construct and induced by 0.2 mM IPTG for protein expression. Cells were resuspended in lysis buffer (50 mM NaH_2PO_4, 300 mM NaCl, 10 mM imidazole pH 8.0) and sonicated. The lysate was centrifuged at 15,000*g* for 20 min to collect supernatant used for SiMPull.

For expression in mammalian cells, YFP–His_6 was generated through addition of a 6×His-tag to the carboxy-terminal of YFP and subcloned into the XhoI and XbaI sites of pCDNA3.1$^+$. A second YFP was subcloned into the HindIII and EcoRI sites of pCDNA3.1-YFP-His_6 to make a tandem dimeric YFP construct. Monomeric and dimeric YFP constructs were transiently expressed in HEK293 cells and purified using standard Ni-NTA chromatography. Proteins were detected by western blot using GFP antibody (Clontech) or Penta-His antibody (Qiagen). For single-molecule analysis, samples were immobilized on biotinylated anti-Penta-His antibody (Qiagen) or on biotinylated polyclonal anti-GFP antibody (Rockland Immunochemicals).

PKA constructs and pull-down. HEK293 cells were transfected with R–Flag–mCherry and C–HA–YFP constructs. The regulatory subunit used was PKA RIIβ, and the catalytic subunit is the Cα isoform. After 24 h expression, cells were harvested into lysis buffer (10 mM Tris pH 7.5, 1% NP-40, 150 mM NaCl, 1 mM EDTA, 1 mM benzamidine, 10 μg ml^{-1} leupeptin, 1 mM NaF, 1 mM Na_3VO_4). This lysate was centrifuged at 14,000*g* for 20 min and used for SiMPull. For bulk immunoprecipitation, anti-Flag M2 beads were added to the lysate for 3 h at 4 °C. Proteins were separated by SDS–PAGE and transferred onto nitrocellulose membranes for blot with anti-HA antibody and anti-mCherry antibody.

For cAMP treatment, a non-hydrolysable analogue (8-Br-cAMP; Sigma) was used to activate PKA. For *in vivo* stimulation, R–Flag–mCherry and C–HA–YFP were transiently expressed in HEK293 cells for 24 h. Cells were pre-treated for 10 min with 10 μM 3-isobutyl-1-methlxanthine (IBMX; Sigma) followed by 5 min stimulation with 10 μM forskolin (Sigma). Cells were immediately washed with cold PBS and lysed as described earlier.

Adrenergic receptor constructs and pull-down. Flag–YFP–β_2AR, HA–YFP–β_2AR, Flag–mCherry–β_1AR and HA–mCherry–β_1AR were transiently expressed in HEK293 cells for 24 h. Cells were harvested into hypotonic lysis buffer (10 mM Tris pH 7.4, 1 mM EDTA, 1 mM benzamidine, 10 μg ml^{-1} leupeptin, 0.3% n-dodecyl-B-D-maltoside (DDM), and incubated for 30 min before centrifugation at 600*g* for 10 min. Supernatants were collected and used for SiMPull with antibodies as indicated.

Mitochondria preparation and MAVS pull-down. HEK293 cells were transiently transfected with YFP–MAVS[30]. Intact mitochondria were isolated using MITOISO2 kit (Sigma) and diluted in the storage buffer supplied with the kit. Mitochondrial preparation was immobilized on slides either directly or after solubilization by adding 1% DDM to the storage buffer. We obtained similar results using whole-cell lysates prepared using several different lysing solutions. YFP–MAVS or mitochondria were immunoprecipitated using biotinylated antibody against GFP.

mTORC1 construct and pull-down. Flag–mTOR was stably transfected in HEK293 cells to obtain near endogenous expression levels of mTOR. For mTORC1 pull-down, Flag–mTOR stable cell lines were transiently transfected with HA–Raptor. HA–Raptor-only lysate was obtained by transiently transfecting HEK293 cells with HA–Raptor. Cells were lysed using CHAPS detergent buffer (40 mM HEPES pH 7.5, 0.3% CHAPS, 150 mM NaCl, 2.5 mM sodium pyrophosphate, 1 mM β-glycerophosphate, 1 mM EDTA), with protease inhibitor cocktail. Lysate was diluted in the buffer without CHAPS for SiMPull with biotinylated anti-Flag antibody (Sigma). Co-immunoprecipitated HA–Raptor was detected using goat anti-HA antibody (Genscript) and donkey anti-goat secondary antibody (Rockland Immunochemicals) labelled with Cy3.

Endogenous protein pull-down. Mouse brain and heart extracts were prepared from 2-week-old FVB mice after anaesthetization. Both samples were homogenized in lysis buffer (20 mM HEPES pH 7.4, 0.5% Triton X-100, 150 mM NaCl, 10% glycerol, 5 mM EDTA, 5 μg ml^{-1} pepstatin, 1 mM PMSF, 1 mM NaF, 1 mM Na_3VO_4), incubated at 4 °C for 1 h followed by centrifugation at 16,000*g* for 10 min to collect the supernatant. These samples were directly used for SiMPull. For bulk immunoprecipitation, protein-A beads were added to pre-clean the lysate (2 h incubation at 4 °C). PKARII antibody was then added to the lysate for overnight immunoprecipitation followed by 1 h incubation with protein-A beads. Control rabbit IgG was added at a final concentration of 2 μg ml^{-1}. The immunoprecipitated proteins were separated by SDS–PAGE and transferred onto nitrocellulose membranes for immunoblot analysis with AKAP150 and PKARII antibodies.

PcrA pull-down and functional assay. His_6-tagged PcrA purified protein and cell lysate were prepared as previously described[35]. The protein was immobilized on slides using antibodies against the polyhistidine tag. A Cy3 and Cy5 dual-labelled partial duplex DNA (Integrated DNA Technologies) with a 5′ tail was added to the immobilized protein. The sequence of DNA used was: 5′ Cy3-$(dT)_{40}$-GCCTCGCTGCCGTCGCCA-3′ + 5′-TGGCGACGGCAGCGAGGC-3′-Cy5.

41. Leake, M. C. *et al.* Variable stoichiometry of the TatA component of the twin-arginine protein transport system observed by *in vivo* single-molecule imaging. *Proc. Natl Acad. Sci. USA* **105,** 15376–15381 (2008).
42. Ulbrich, M. H. & Isacoff, E. Y. Rules of engagement for NMDA receptor subunits. *Proc. Natl Acad. Sci. USA* **105,** 14163–14168 (2008).

LETTER

doi:10.1038/nature10077

Hf–W–Th evidence for rapid growth of Mars and its status as a planetary embryo

N. Dauphas[1] & A. Pourmand[1,2]

Terrestrial planets are thought to have formed through collisions between large planetary embryos[1] of diameter ~1,000–5,000 km. For Earth, the last of these collisions involved an impact by a Mars-size embryo that formed the Moon 50–150 million years (Myr) after the birth of the Solar System[2,3]. Although model simulations of the growth of terrestrial planets can reproduce the mass and dynamical parameters of the Earth and Venus, they fall short of explaining the small size of Mars[4,5]. One possibility is that Mars was a planetary embryo that escaped collision and merging with other embryos[1]. To assess this idea, it is crucial to know Mars' accretion timescale[6], which can be investigated using the ^{182}Hf–^{182}W decay system in shergottite-nakhlite-chassignite meteorites[6–10]. Nevertheless, this timescale remains poorly constrained owing to a large uncertainty associated with the Hf/W ratio of the Martian mantle[6] and as a result, contradicting timescales have been reported that range between 0 and 15 Myr (refs 6–10). Here we show that Mars accreted very rapidly and reached about half of its present size in only $1.8^{+0.9}_{-1.0}$ Myr or less, which is consistent with a stranded planetary embryo origin. We have found a well-defined correlation between the Th/Hf and ^{176}Hf/^{177}Hf ratios in chondrites that reflects remobilization of Lu and Th during parent-body processes. Using this relationship, we estimate the Hf/W ratio in Mars' mantle to be 3.51 ± 0.45. This value is much more precise than previous estimates, which ranged between 2.6 and 5.0 (ref. 6), and lifts the large uncertainty that plagued previous estimates of the age of Mars. Our results also demonstrate that Mars grew before dissipation of the nebular gas when ~100-km planetesimals, such as the parent bodies of chondrites, were still being formed. Mars' accretion occurred early enough to allow establishment of a magma ocean powered by decay of ^{26}Al.

The age of core formation on Mars can be estimated by measuring the excess abundance of ^{182}W, which is produced from decay of ^{182}Hf, relative to other non-radiogenic isotopes of W in the Martian mantle. Measurements of shergottite–nakhlite–chassignite (SNC) meteorites give a value of $\varepsilon^{182}W_{\text{Mars mantle}} \approx +0.4$ or possibly higher (here ε^{182}W is the deviation in 0.01% of the ^{182}W/^{183}W or ^{182}W/^{184}W ratio relative to the terrestrial mantle)[6–10]. In addition to ε^{182}W, the Hf/W ratio of Mars' mantle is also needed to calculate the age of core formation. This ratio, however, cannot be measured directly in SNC meteorites because Hf and W behave differently during mantle melting and crystallization. Instead, the mantle Hf/W ratio can be calculated as $(Hf/W)_{\text{Mars mantle}} = (Th/W)_{\text{Mars mantle}}/(Th/Hf)_{\text{Mars mantle}}$. In most circumstances, Th and W have similar magmatic behaviours[6]. Despite different sources and chemical compositions, the average Th/W ratios of shergottites and nakhlites are identical within uncertainties (that is, 0.68 ± 0.09 versus 0.79 ± 0.32, Supplementary Table 1). This shows that there is no relationship between measured Th/W ratios and the inferred mineralogy of the mantle source region (for example, the presence of clinopyrexene, garnet or ilmenite[11,12]). There is also no relationship between Th/W ratios and geochemical proxies of magmatic fractionation[6]. Therefore, $(Th/W)_{\text{Mars mantle}}$ can be approximated by the average value of SNC meteorites, $(Th/W)_{\text{Mars mantle}} = (Th/W)_{\text{SNC}} = 0.752 \pm 0.092$ (Supplementary Table 1). As refractory lithophile elements, Th and Hf are expected to be in chondritic proportions (CHUR, chondritic uniform reservoir) in the Martian mantle, hence, $(Th/Hf)_{\text{Mars mantle}} = (Th/Hf)_{\text{CHUR}}$. It follows that:

$$(Hf/W)_{\text{Mars mantle}} = (Th/W)_{\text{SNC}}/(Th/Hf)_{\text{CHUR}} \quad (1)$$

The main source of uncertainty in the age of Mars is the Th/Hf ratio in chondrites that varies by a factor of >3 (ref. 6, this study). In order to address this problem, we have measured elemental concentrations (U, Th, Lu and Hf) and isotopic compositions (^{176}Hf/^{177}Hf) of 44 chondrites, mainly falls, covering all major chondrite groups (Table 1). The elemental concentrations were measured by isotope dilution; the methodology is described in Methods Summary. The meteorites define a clear correlation in Th/Hf versus ^{176}Hf/^{177}Hf space (Fig. 1). The ^{176}Hf/^{177}Hf ratio is another proxy for Lu/Hf ratio, as ^{176}Lu decays to ^{176}Hf with a half-life of 36.8 Gyr (refs 13, 14). The variations in Hf isotopic composition of ordinary chondrites have been interpreted to reflect redistribution of Lu during parent-body metamorphism[14]. Lu, U and Th are concentrated in trace carrier phases like phosphates (chlorapatite and merrillite) for ordinary as well as carbonaceous chondrites, and sulphides (oldhamite) for enstatite chondrites (refs 15, 16 and references therein). Variations in the abundance of these trace phases at the bulk sample scale due to parent-body redistribution during metamorphism (that is, a nugget effect; Fig. 1b) is the most likely cause for the observed correlation in Fig. 1a. The chondritic ^{176}Hf/^{177}Hf isotopic ratio (~0.282785) is well known from measurement of chondrites of low metamorphic grade[14]. We can therefore use this ratio and the correlation presented in Fig. 1a to calculate $(Th/Hf)_{\text{Mars mantle}} = (Th/Hf)_{\text{CHUR}} = 0.2144 \pm 0.0075$. This translates to a precise estimate of 3.51 ± 0.45 for the Hf/W ratio of the Martian mantle (equation (1)).

If Mars was a stranded embryo formed by accretion of planetesimals during runaway and oligarchic growth (see below), its mass at any given time can be parameterized as[17,18]

$$M_{\text{Mars}}(t)/M_{\text{Mars}} = \tanh^3(t/\tau) \quad (2)$$

where t is counted from the formation of the Solar System (that is, the condensation of calcium-aluminium-rich inclusions, CAIs, in meteorites), and τ is the accretion timescale. At $t = \tau$, the embryo reached $\tanh^3(1) = 44\%$ of its present size. On Earth, the extent to which 1,000–5,000-km planetary embryos striking the surface equilibrated with the mantle of the protoplanet is uncertain. Modelling and geochemical considerations, however, suggest that incomplete equilibration on Earth affected Hf–W chronology[19,20]. If Mars were an embryo, its accretion would have proceeded by collisions with 10–100-km planetesimals that were much smaller than the target embryo. In addition, Mars would have formed early enough to develop a magma ocean from the heat released by ^{26}Al decay (<2.5 Myr after

[1]Origins Laboratory, Department of the Geophysical Sciences and Enrico Fermi Institute, The University of Chicago, 5734 South Ellis Avenue, Chicago, Illinois 60637, USA. [2]Rosenstiel School of Marine & Atmospheric Science, Division of Marine Geology and Geophysics, University of Miami, 4600 Rickenbacker Causeway, Miami, Florida 33149, USA.

Table 1 | U–Th–Lu–Hf systematics of chondrites

Meteorite name	Meteorite group	Fall	Source	Collection ID	U ($ng\,g^{-1}$)	Th ($ng\,g^{-1}$)	Hf ($ng\,g^{-1}$)	Lu ($ng\,g^{-1}$)	$^{176}Hf/^{177}Hf$ (JMC-normalized)	Lu/Hf atomic ratio	Th/Hf atomic ratio	Th/U atomic ratio	Lu/Th atomic ratio
Ivuna	CI1	Yes	USNM	6630	8.29	31.85	116.22	27.99	0.282826	0.2457	0.2108	3.94	1.166
Mighei	CM2	Yes	FM	1456	10.64	37.69	139.05	33.01	0.282795	0.2421	0.2085	3.63	1.161
Kainsaz	CO3.2	Yes	FM	2755	13.73	48.00	186.17	43.75	0.282817	0.2397	0.1983	3.59	1.209
Lancé	CO3.5	Yes	FM	1351	16.18	47.27	177.31	41.46	0.282762	0.2385	0.2051	3.00	1.163
Allende	CV3	Yes	USNM	3529	15.34	58.19	192.17	46.02	0.282835	0.2443	0.2329	3.89	1.049
Allende	CV3	Yes	USNM	3529	16.00	58.04	192.16	45.92	0.282795	0.2438	0.2323	3.72	1.049
Allende	CV3	Yes	USNM	3529	15.71	58.90	189.49	45.24	0.282839	0.2436	0.2391	3.85	1.019
Allende	CV3	Yes	USNM	3529	15.38	59.75	205.55	48.54	0.282814	0.2409	0.2236	3.98	1.077
Allende*	CV3	Yes	USNM	3529	15.58	59.80	192.93	46.03	0.282824	0.2434	0.2384	3.94	1.021
Allende*	CV3	Yes	USNM	3529	15.23	58.37	192.70	46.43	0.282795	0.2458	0.2330	3.93	1.055
Allende*	CV3	Yes	USNM	3529	15.57	57.42	192.20	46.12	0.282801	0.2448	0.2298	3.78	1.065
Grosnaja	CV3	Yes	FM	1732	23.61	61.50	161.60	39.07	0.282811	0.2466	0.2927	2.67	0.843
Vigarano	CV3	Yes	FM	782	14.66	53.49	207.07	49.15	0.282789	0.2421	0.1987	3.74	1.219
Sahara 97072	EH3	No	Private collection	NA	8.92	28.40	111.02	25.27	0.282748	0.2322	0.1968	3.27	1.180
Qingzhen	EH3	Yes	R.N. Clayton	NA	7.61	25.84	107.09	24.50	0.282704	0.2334	0.1856	3.48	1.257
Indarch	EH4	Yes	FM	3466	9.44	27.37	104.49	23.45	0.282685	0.2290	0.2015	2.97	1.136
Adhi Kot	EH4	Yes	AMNH	3993	7.86	26.76	73.31	23.92	0.284055	0.3329	0.2808	3.49	1.186
St. Mark's	EH5	Yes	USNM	3027	8.39	27.60	87.68	23.18	0.283143	0.2697	0.2421	3.38	1.114
Saint-Sauveur	EH5	Yes	MNHN	1456	7.62	24.40	76.07	20.31	0.283250	0.2724	0.2468	3.29	1.104
Daniel's Kuil	EL6	Yes	FM	1500	8.14	29.34	138.16	29.26	0.282528	0.2161	0.1633	3.70	1.323
Yilmia	EL6	No	FM	2740	5.53	23.05	117.58	23.41	0.282331	0.2031	0.1508	4.28	1.347
Blithfield	EL6	No	FM	1646	18.75	48.31	103.66	34.24	0.284231	0.3369	0.3585	2.64	0.940
Hvittis	EL6	Yes	FM	578	6.82	31.89	134.41	30.38	0.282651	0.2306	0.1825	4.80	1.263
Eagle	EL6	Yes	FM	3149	7.56	31.05	176.23	31.20	0.281972	0.1806	0.1356	4.22	1.332
Khairpur	EL6	Yes	FM	1538	6.84	25.77	137.20	24.05	0.282061	0.1788	0.1445	3.86	1.237
Pillistfer	EL6	Yes	FM	1647	5.82	25.16	112.72	24.97	0.282648	0.2260	0.1717	4.43	1.316
Jajh Deh Kot Lalu	EL6	Yes	USNM	1260	4.90	24.27	158.95	22.88	0.281675	0.1468	0.1175	5.08	1.250
Happy Canyon	EL6/7	No	FM	2760	189.63	30.47	141.55	23.12	0.281758	0.1666	0.1656	0.16	1.006
Ilafegh 009	EL7	No	AMNH	4757	5.48	26.19	100.26	24.04	0.282726	0.2446	0.2009	4.90	1.218
Bielokrynitschie	H4	Yes	FM	1394	12.17	40.79	144.31	31.37	0.282581	0.2218	0.2174	3.44	1.020
Ochansk	H4	Yes	FM	1443	11.63	39.46	151.14	33.96	0.282699	0.2292	0.2008	3.48	1.141
Kesen	H4	Yes	FM	1822	12.71	38.79	132.67	32.36	0.282849	0.2489	0.2249	3.13	1.106
Kernouve	H6	Yes	MNHN	602	12.03	35.68	128.88	29.30	0.282695	0.2319	0.2130	3.04	1.089
Dalgety Downs	L4	No	FM	2613	21.87	49.11	144.68	31.79	0.282371	0.2242	0.2611	2.30	0.859
Bald Mountain	L4	Yes	FM	2392	10.45	38.36	147.24	37.14	0.283005	0.2573	0.2004	3.77	1.284
Barratta	L4	No	FM	1463	13.50	38.71	148.75	34.66	0.282760	0.2377	0.2002	2.94	1.187
Farmington	L5	Yes	FM	347	20.73	53.82	148.21	44.17	0.283741	0.3040	0.2793	2.66	1.088
Harleton	L6	Yes	FM	2686	8.83	34.56	150.25	31.96	0.282421	0.2170	0.1769	4.01	1.226
Hamlet	LL4	Yes	FM	3296	12.98	47.57	177.02	39.01	0.282619	0.2248	0.2067	3.76	1.087
Kelly	LL4	No	FM	2235	15.07	38.50	119.33	32.18	0.283067	0.2751	0.2481	2.62	1.109
Soko-Banja	LL4	Yes	FM	1374	12.42	39.01	149.07	33.83	0.282722	0.2315	0.2013	3.22	1.150
Tuxtuac	LL5	Yes	FM	2850	18.01	50.03	168.49	38.91	0.282619	0.2356	0.2284	2.85	1.032
Paragould	LL5	Yes	FM	2135	21.61	58.41	171.60	45.16	0.283217	0.2684	0.2618	2.77	1.025
Saint-Séverin	LL6	Yes	MNHN	2397	11.16	47.64	184.68	42.70	0.282735	0.2358	0.1984	4.38	1.189

Concentration measurements were carried out by isotope dilution mass spectrometry. The absolute uncertainties (2σ, 95% confidence intervals calculated from replicate analyses of Allende) are ±0.52 (U), ±1.78 (Th), ±10.54 (Hf), ±2.08 (Lu), ±0.000036 ($^{176}Hf/^{177}Hf$), ±0.0030 (Lu/Hf), ±0.0105 (Th/Hf), ±0.19 (Th/U) and ±0.043 (Lu/Th). 'JMC-normalized' refers to JMC-475 (see Methods for details). Happy Canyon has a high U/Th ratio, reflecting terrestrial contamination, and was not used in calculation of the Th/Hf ratio of CHUR. NA, not available.
* Samples were digested by high-pressure HF bomb. All other samples were digested by flux fusion.

CAIs, ref. 21), and the planetesimals striking its surface would have also been molten. The fact that the impactors were molten and that they were much smaller than the target embryo would have led to vaporization of the planetesimals upon impact and equilibration with the mantle of proto-Mars. Under this reasonable assumption, we can predict the excess in radiogenic ^{182}W of the mantle relative to chondrites for various accretion timescales (values of τ)[9]:

$$\varepsilon^{182}W_{\text{Mars mantle}} - \varepsilon^{182}W_{\text{CHUR}} = q_W \left(\frac{^{182}Hf}{^{180}Hf}\right)_0 f^{Hf/W}_{\text{Mars mantle}} \lambda \int_0^{4{,}568\text{Myr}} e^{-\lambda t} \tanh^{3+3f^{Hf/W}_{\text{Mars mantle}}}(t/\tau)\, dt \quad (3)$$

All parameters except $f^{Hf/W}_{\text{Mars mantle}} = (Hf/W)_{\text{Mars mantle}}/(Hf/W)_{\text{CHUR}} - 1 = 3.38 \pm 0.56$ were known before this work; $\varepsilon^{182}W_{\text{CHUR}} = -2.23 \pm 0.11$, $q_W = (^{180}Hf/^{182}W)_{\text{CHUR}} \times 10^4 = 1.07 \times 10^4$, $(^{182}Hf/^{180}Hf)_0 = (9.72 \pm 0.44) \times 10^{-5}$, and λ, the decay constant of ^{182}Hf, is $0.0779 \pm 0.0008\ Myr^{-1}$ (refs 7, 10, 22; see Supplementary Information for details). A mixture of 85% H, 11% CV and 4% CI chondrites is chosen for CHUR, which was previously proposed as a model composition for bulk Mars[23]. These mixing proportions influence the values q_W, $\varepsilon^{182}W_{\text{CHUR}}$ and Hf/W_{CHUR}. The W isotopic composition of the Martian mantle is uncertain because different Martian meteorites show variable $\varepsilon^{182}W$ values (from +0.3 to +3). This reflects fractionation of the Hf/W ratio in the Martian magma ocean while ^{182}Hf was still present[7–10].

We adopt a conservative approach by using the least radiogenic W isotopic compositions (lowest $\varepsilon^{182}W$ value) of $\varepsilon^{182}W_{\text{Mars mantle}} = +0.45 \pm 0.15$ (refs 8, 10) proposed for the bulk Martian mantle. Higher values such as those documented in nakhlites would translate to a shorter accretion time. Therefore, the timescale that we calculate is a robust upper limit. To reproduce $\varepsilon^{182}W_{\text{Mars mantle}} = +0.45 \pm 0.15$, we calculate $\tau = 1.8^{+0.9}_{-1.0}$ Myr (Fig. 2), where the error bar (yellow band) represents the 95% confidence interval and is obtained by propagating the uncertainties on $\varepsilon^{182}W_{\text{Mars mantle}}$, $\varepsilon^{182}W_{\text{CHUR}}$, $(^{182}Hf/^{180}Hf)_0$, $f^{Hf/W}_{\text{Mars mantle}}$ and λ, using a Monte Carlo simulation. Assuming a different composition for CHUR would not change the results significantly (for example, $\tau = 2.0^{+1.0}_{-1.2}$ Myr for 100% CI). Scenarios that invoke late-stage accretion of a large body to explain Mars' hemispheric dichotomy[24] could increase model ages in the case of incomplete W isotope equilibration between the impactor and the Martian mantle.

As seen in Fig. 2, most of the accretion of Mars took place during the first 4 Myr of the formation of the Solar System. The accretion

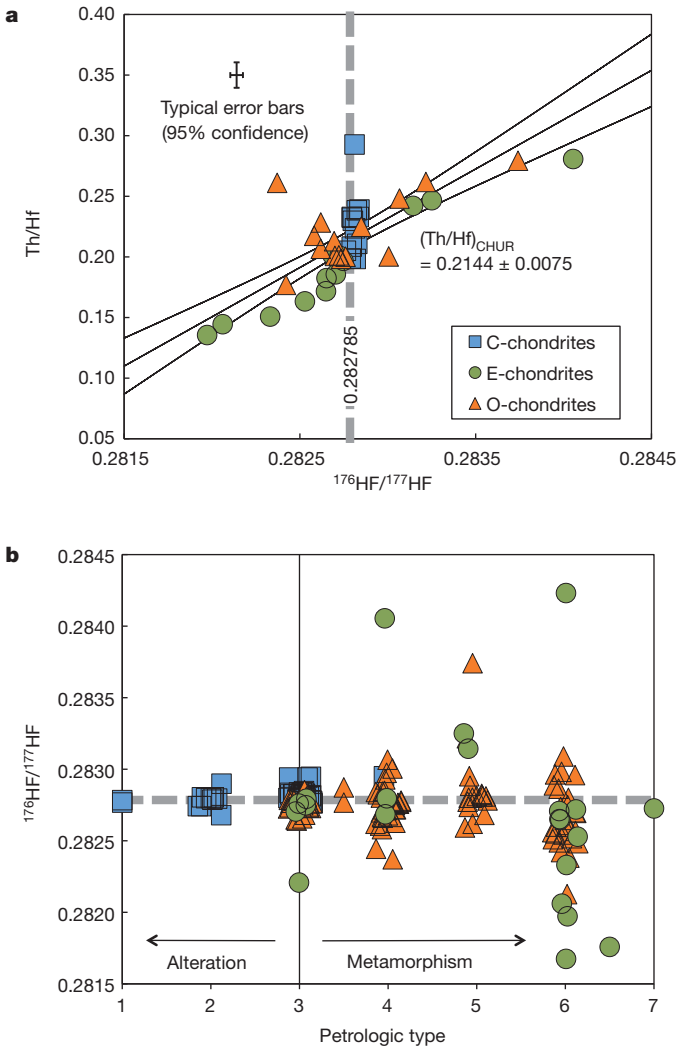


Figure 1 | Determination of the Th/Hf ratio of CHUR. **a**, Correlation between Th/Hf and $^{176}Hf/^{177}Hf$ (a proxy for Lu/Hf) ratios in chondrites (Table 1; C, carbonaceous; E, enstatite; O, ordinary; meteorite Happy Canyon not included). The $^{176}Hf/^{177}Hf$ ratio of CHUR is estimated to be 0.282785 ± 0.000011 (ref. 14), allowing us to estimate $(Th/Hf)_{CHUR} = 0.2144 \pm 0.0075$ (the uncertainty is the 95% confidence interval based on regression of the data). At a Th/Hf ratio of 0, we calculate a $^{176}Hf/^{177}Hf$ ratio of ~0.280, which is close to the Solar System initial ratio of 0.27978 (ref. 14). There is no correlation between the Lu/Th and Lu/Hf ratios (Table 1). These two observations suggest that Th behaves very similarly to Lu but that both elements can be decoupled from Hf in meteorites. If chondrite parent bodies had begun with different Th/Hf ratios (for example, owing to evaporation/condensation processes in the nebula), they would not define a single correlation line. **b**, $^{176}Hf/^{177}Hf$ ratios of chondrites as a function of petrologic types (data from Table 1, refs 12, 13 and references therein). The data points have been moved horizontally by random values to decrease overlap and improve readability. The degree of aqueous alteration increases from type 3 to 1, while the degree of thermal metamorphism increases from type 3 to 7 (that is, the most pristine samples are of type 3). The dispersion in $^{176}Hf/^{177}Hf$ ratios of metamorphosed chondrites (types 4–7) is much larger compared with other samples, indicating that redistribution during parent-body metamorphism is the most likely explanation for the dispersion in Lu/Hf (and Th/Hf) ratios measured in bulk chondrite specimens. The average Th/Hf ratio of unmetamorphosed chondrites (types 1–3) is 0.2217 ± 0.0135, identical within uncertainties with the value derived from the regression in **a**.

timescale of Mars can be compared with those of planetesimals, whose formation history is known from measurement of meteorites. The parent bodies of iron meteorites were formed at the same time or soon after CAIs[10]. Chondrites, which formed >3 Myr after CAIs[25], contain pristine materials such as CAIs and presolar grains that indicate they

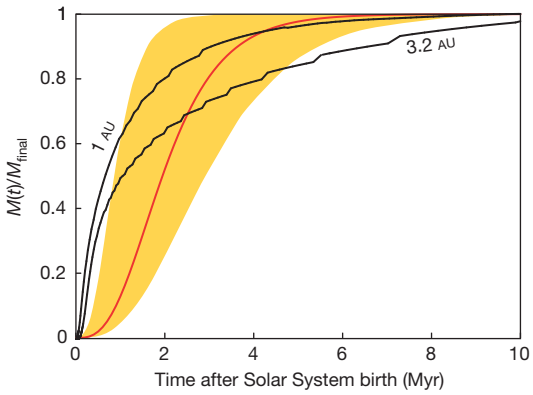


Figure 2 | Accretion timescale of Mars inferred from ^{182}Hf–^{182}W systematics. The red curve is calculated from $M(t)/M_{final} = \tanh^3(t/\tau)$ with $\tau = 1.8^{+0.9}_{-1.0}$ Myr (see text for details). The yellow band corresponds to the 95% confidence interval on the accretion curve of Mars obtained by propagating all uncertainties on model parameters using a Monte Carlo simulation. The two black curves are model simulations of embryo growth at 1 and 3.2 AU (ref. 26) for comparison.

accreted from pristine nebular dust, rather than from collisional debris. Detailed thermochronology also shows that chondrites' parent bodies were probably around 100 km in size[21]. An important finding of this work, therefore, is that large planetesimals (for example, 10–100-km bodies) were still being formed in the protoplanetary disk during Mars' accretion.

The accretion of planetary embryos from planetesimals began with a period of runaway growth, in which large bodies accreted at a much higher rate than smaller ones. When the larger bodies (that is, embryos) became massive enough to dynamically stir the planetesimals around them, their growth was slowed down by reduced gravitational focusing factors, a process known as 'oligarchic growth'. The embryos eventually cleared their orbits of smaller objects, limiting their mass to Moon- to Mars-sized objects. Computer simulations predict that the growth of embryos at 1.5 astronomical units (AU) from the Sun should occur on a maximal timescale of a few million years (note that this timescale is sensitive to the assumed initial surface density)[17,18,26]. The growth of terrestrial planets such as Earth, on the other hand, is thought to have proceeded by collisions between these embryos on an expected timescale of several tens of millions of years, a process known as 'chaotic growth'[1,5]. This timescale on Earth is best constrained by the age of the Moon, which is thought to have formed by a collision that occurred 50–150 Myr after CAIs[3] between a Mars-size planetary embryo and the proto-Earth[2]. It is difficult to reconcile Mars' accretion timescale based on the $\varepsilon^{182}W$ data ($\tau = 1.8^{+0.9}_{-1.0}$ Myr) with an Earth-like mode of accretion. Our findings, however, are entirely consistent with the timescale predicted by models of oligarchic growth (Fig. 2). We therefore conclude that Mars is most likely to be an embryo that escaped collisions and merging with other bodies, thus explaining its small mass compared to Earth and Venus.

Rapid accretion of Mars also has important implications for its magmatic history. The gravitational energy deposited on the growing embryo from planetesimals during oligarchic growth would not have been sufficient to cause large-scale melting[27], except for a mode of runaway core formation triggered by impacts with >3,500-km embryos[28]. In this scheme, the temperature rise would not have been sufficient to induce silicate melting, and molten metal would have segregated through a solid matrix. The Martian mantle (based on the composition of SNC meteorites) contains large ^{182}W and ^{142}Nd isotopic heterogeneities that arise from magmatic fractionation of Hf/W and Sm/Nd while ^{182}Hf and ^{146}Sm (half-life 103 Myr) were still present[7,8,29,30]. The event that caused this chemical fractionation is dated at ~20–60 Myr after CAIs, and may correspond to crystallization of a magma ocean on Mars. A similar age range is obtained from

^{129}I–^{244}Pu–Xe systematics of SNC meteorites, indicating rapid degassing of the Martian mantle[31]. Thermal modelling shows that planetary bodies accreted before ~2.5 Myr would have incorporated enough of the short-lived nuclide ^{26}Al (half-life 0.717 Myr) for radioactive decay (3.16 MeV per decay) to induce silicate melting[21]. Mars would have reached 69±30% of its present size by that time and the heat generated from ^{26}Al decay alone would have been sufficient to establish a magma ocean.

Finally, the idea that Mars was a planetary embryo may also explain the similarities between some characteristics of the terrestrial and Martian atmospheres in spite of marked differences in their sizes and accretion histories. The atmospheres of both planets share what is commonly referred to as the 'missing xenon problem'; on both planets, Xe is isotopically fractionated by 3–4% per AMU relative to the solar value and to that in chondrites, the Xe/Kr ratio is close to the solar value, and Kr is only slightly fractionated[32]. If hydrodynamic escape were responsible for isotopic fractionation of noble gases in the terrestrial planet atmospheres, one would expect more depletion and higher isotopic fractionation in lighter Kr compared with Xe, which is the opposite to what is actually observed. In some models that attempt to explain this problem quantitatively, the similarity between Mars and Earth is taken as a coincidence[32,33]. Alternatively, if Mars was a planetary embryo, as suggested by our findings, Earth may have inherited its missing Xe problem from the atmosphere of a Mars-like planetary embryo, possibly the impactor that also formed the Moon. This idea is consistent with the time when Earth became retentive for Xe, which is estimated to be ~100 Myr after the birth of the Solar System and may correspond to the time of the Moon-forming giant impact[19].

METHODS SUMMARY

About 100 mg of homogenized powder was fused with ultra-pure lithium metaborate (+LiBr) in high-purity graphite crucibles at 1,070 °C for 12 min. The fusion melt was dissolved and spiked with a calibrated ^{236}U–^{229}Th–^{176}Lu–^{180}Hf solution for isotope dilution mass spectrometry (IDMS). After sample–spike equilibration and separation of matrix elements on a 2-ml Eichrom TODGA cartridge, Hf, U and Th were eluted in 3 mol l^{-1} HNO_3 + 0.3 mol l^{-1} HF at 65 °C. Lu was eluted in 0.5 mol l^{-1} HCl and was further purified from heavy rare earth elements, using a 2-ml Eichrom Ln cartridge. The solutions were evaporated and the residues were redissolved in 0.4 mol l^{-1} HNO_3 + 0.04 mol l^{-1} HF for Hf, U and Th, and 0.4 mol l^{-1} HNO_3 for Lu. About 90% of the Hf–U–Th solution was used for concentration and high-precision isotope analysis of Hf, while 10% was analysed for U and Th concentrations by IDMS. All the measurements were done on a Neptune multi-collector inductively coupled plasma mass spectrometer at the University of Chicago. Meteorites and geostandards of known Hf isotopic and U–Th–Lu–Hf elemental compositions were measured to validate the method. Uncertainties (2σ, 95% confidence intervals) were calculated on the basis of replicate analyses of the Allende CV3 chondrite.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 20 October 2010; accepted 22 March 2011.

1. Chambers, J. E. & Wetherill, G. W. Making the terrestrial planets: *N*-body integrations of planetary embryos in three dimensions. *Icarus* **136,** 304–327 (1998).
2. Canup, R. M. & Asphaug, E. Origin of the Moon in a giant impact near the end of the Earth's formation. *Nature* **412,** 708–712 (2001).
3. Touboul, M., Kleine, T., Bourdon, B., Palme, H. & Wieler, R. Late formation and prolonged differentiation of the Moon inferred from W isotopes in lunar metals. *Nature* **450,** 1206–1209 (2007).
4. Wetherill, G. W. Why isn't Mars as big as Earth? *Lunar Planet. Sci.* **XXII,** 1495–1496 (1991).
5. Raymond, S. N., O'Brien, D. P., Morbidelli, A. & Kaib, A. Building the terrestrial planets: constrained accretion in the inner solar system. *Icarus* **203,** 644–662 (2009).
6. Nimmo, F. & Kleine, T. How rapidly did Mars accrete? Uncertainties in the Hf-W timing of core formation. *Icarus* **191,** 497–504 (2007).
7. Kleine, T., Mezger, K., Münker, C., Palme, H. & Bischoff, A. ^{182}Hf-^{182}W isotope systematics of chondrites, eucrites, and martian meteorites: chronology of core formation and early mantle differentiation in Vesta and Mars. *Geochim. Cosmochim. Acta* **68,** 2935–2946 (2004).
8. Foley, C. N. *et al.* The early differentiation history of Mars from ^{182}W-^{142}Nd isotope systematics in the SNC meteorites. *Geochim. Cosmochim. Acta* **69,** 4557–4571 (2005).
9. Jacobsen, S. B. The Hf-W isotopic system and the origin of the Earth and Moon. *Annu. Rev. Earth Planet. Sci.* **33,** 531–570 (2005).
10. Kleine, T. *et al.* Hf-W chronology of the accretion and early evolution of asteroids and terrestrial planets. *Geochim. Cosmochim. Acta* **73,** 5150–5188 (2009).
11. Righter, K. & Shearer, C. K. Magmatic fractionation of Hf and W: constraints on the timing of core formation and differentiation in the Moon and Mars. *Geochim. Cosmochim. Acta* **67,** 2497–2507 (2003).
12. Bouvier, A., Blichert-Toft, J. & Albarède, F. Martian meteorite chronology and evolution of the interior of Mars. *Earth Planet. Sci. Lett.* **280,** 285–295 (2009).
13. Blichert-Toft, J. & Albarède, F. The Lu-Hf isotope geochemistry of chondrites and the evolution of the mantle-crust system. *Earth Planet. Sci. Lett.* **148,** 243–258 (1997).
14. Bouvier, A., Vervoort, J. D. & Patchett, P. J. The Lu-Hf and Sm-Nd isotopic composition of CHUR: constraints from unequilibrated chondrites and implications for the bulk composition of terrestrial planets. *Earth Planet. Sci. Lett.* **273,** 48–57 (2008).
15. Goreva, J. S. & Burnett, D. S. Phosphate control on the thorium/uranium variations in ordinary chondrites: improving solar system abundances. *Meteorit. Planet. Sci.* **36,** 63–74 (2001).
16. Murrell, M. T. & Burnett, D. S. Actinide microdistributions in the enstatite meteorites. *Geochim. Cosmochim. Acta* **46,** 2453–2460 (1982).
17. Thommes, E. W., Duncan, M. J. & Levison, H. F. in *Astrophysical Ages and Time Scales* (eds von Hippel, T., Simpson, C. & Manset, N.) 91–100 (ASP Conference Series Vol. 245, Astronomical Society of the Pacific, 2001).
18. Chambers, J. A semi-analytic model for oligarchic growth. *Icarus* **180,** 496–513 (2006).
19. Halliday, A. N. Mixing, volatile loss and compositional change during impact-driven accretion of the Earth. *Nature* **427,** 505–509 (2004).
20. Dahl, T. W. & Stevenson, D. J. Turbulent mixing of metal and silicate during planetary accretion — an interpretation of the Hf-W chronometer. *Earth Planet. Sci. Lett.* **295,** 177–186 (2010).
21. Grimm, R. E. & McSween, H. Y. Jr. Heliocentric zoning of the asteroid belt by aluminum-26 heating. *Science* **259,** 653–655 (1993).
22. Yin, Q. *et al.* A short timescale for terrestrial planet formation from Hf-W chronometry of meteorites. *Nature* **418,** 949–952 (2002).
23. Lodders, K. & Fegley, B. Jr. An oxygen isotope model for the composition of Mars. *Icarus* **126,** 373–394 (1997).
24. Wilhelms, D. E. & Squyres, S. W. The martian hemispheric dichotomy may be due to a giant impact. *Nature* **309,** 138–140 (1984).
25. Kita, N. T. *et al.* in *Chondrites and the Protoplanetary Disk* (eds Krot, A. N., Scott, E. R. D. & Reipurth, B.) 558–587 (ASP Conference Series Vol. 341, Astronomical Society of the Pacific, 2005).
26. Kobayashi, H., Tanaka, H., Krivov, A. V. & Inaba, S. Planetary growth with collisional fragmentation and gas drag. *Icarus* **209,** 836–847 (2010).
27. Senshu, H., Kuramoto, K. & Matsui, T. Thermal evolution of a growing Mars. *J. Geophys. Res.* **107,** 5118, doi:10.1029/2001JE001819 (2002).
28. Ricard, Y., Srámek, O. & Dubuffet, F. A multi-phase model of runaway core-mantle segregation in planetary embryos. *Earth Planet. Sci. Lett.* **284,** 144–150 (2009).
29. Caro, G., Bourdon, B., Halliday, A. N. & Quitté, G. Super-chondritic Sm/Nd ratios in Mars, the Earth and the Moon. *Nature* **452,** 336–339 (2008).
30. Debaille, V., Brandon, A. D., Yin, Q. Z. & Jacobsen, B. Coupled ^{142}Nd-^{143}Nd evidence for a protracted magma ocean in Mars. *Nature* **450,** 525–528 (2007).
31. Marty, B. & Marti, K. Signatures of early differentiation of Mars. *Earth Planet. Sci. Lett.* **196,** 251–263 (2002).
32. Pepin, R. O. On the origin and early evolution of terrestrial planet atmospheres and meteoritic volatiles. *Icarus* **92,** 2–79 (1991).
33. Dauphas, N. The dual origin of the terrestrial atmosphere. *Icarus* **165,** 326–339 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Discussions with M. Chaussidon, H. Kobayashi, F. J. Ciesla, D. J. Stevenson, T. W. Dahl, R. Yokochi and G. Coutrot were appreciated. Comments from A. D. Brandon helped to improve the quality of the manuscript. We thank H. Kobayashi for sharing the digital outputs of his model simulations with us. The meteorite samples were provided by the Field Museum, the Smithsonian, the Muséum National d'Histoire Naturelle and R. N. Clayton. F. Marcantonio and P. J. Patchett gave us solutions of standards and spikes that were used to calibrate the measurements. This work was supported by a Packard fellowship, NASA and the NSF through grants NNX09AG59G and EAR-0820807 to N.D.

Author Contributions Both authors contributed equally to this work. N.D. and A.P. devised the method for purification and analysis of U, Th, Lu and Hf; A.P. performed the meteorite measurements; N.D. did the modelling; N.D. and A.P. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to N.D. (dauphas@uchicago.edu).

METHODS

Details of sample digestion, elemental separations by TODGA extraction chromatography and multi-collector inductively coupled plasma mass spectrometer (MC-ICPMS) analysis have been published elsewhere[34]. Relatively unaltered meteorite chips of 100–500 mg (mainly observed falls) were sawn from larger pieces and cleaned for 10–20 s in an ultrasonic bath with a solution of 1 mol l^{-1} HCl (except for CI and E chondrites) followed by 10 s of sonication in high-purity ethanol. Dried sample pieces were crushed in an agate mortar and ~100 mg of the homogenized powder was directly weighed with ultra-pure lithium metaborate flux in a 8-ml high-purity graphite crucible (Spex CertiPrep) for alkali flux fusion. Flux:sample ratios ≥6 were required in order to achieve complete sample digestion and subsequent dissolution of the fusion melt. High-temperature flux fusion was preferred over acid dissolution to ensure that all refractory phases were completely digested. In order to achieve low-blank dissolution by flux melting, commercially available Puratronic lithium metaborate (99.997% metals basis, Alfa Aesar) was further purified according to the protocol presented in ref. 34. Approximately 100 ml of high-purity lithium bromide (LiBr) solution was also added as non-wetting agent to prevent adhesion of the flux to the graphite crucible and allow quantitative transfer of the melt.

The crucible was capped to minimize contamination and fusion was performed in a Thermolyne furnace at 1,070 °C for 12 min. The melt was transferred to a 30-ml Teflon PFA Savillex vial containing 25 ml of 3 mol l^{-1} HNO_3 and complete dissolution was typically achieved in a few minutes. The flux solution was spiked with a calibrated ^{236}U–^{229}Th–^{176}Lu–^{180}Hf spike solution for IDMS. After 4 h of sample–spike equilibration at 120 °C, the solution was directly loaded onto a pre-conditioned, 2-ml Eichrom TODGA cartridge for separation of the analytes from matrix elements by extraction chromatography. The TODGA resin has a very high affinity for actinides (U and Th), high-field strength elements (Hf) and rare earth elements (Lu) in 3 mol l^{-1} HNO_3 (refs 34, 35). Matrix elements were rinsed from the TODGA cartridge in 12 ml of 3 mol l^{-1} HNO_3, followed by 15 ml of 11 mol l^{-1} HNO_3. Subsequently, U, Th and Hf were eluted in 20 ml of 3 mol l^{-1} HNO_3 + 0.3 mol l^{-1} HF at 65 °C. Ytterbium and Lu were directly eluted onto a 2-ml Eichrom Ln cartridge in 0.5 mol l^{-1} HCl. The additional Ln chromatography step was needed to separate residual heavy rare earth elements (HREE) from Lu and Yb. Following the removal of HREE in 3.5 mol l^{-1} HCl, ~70% of Lu and ~30% of Yb were eluted in 20 ml of 6 mol l^{-1} HCl. The eluents, containing U–Th–Hf and Lu–Yb, were evaporated to near dryness and treated with a few drops of concentrated perchloric acid to eliminate potential organic molecules from the extraction resins. The residues were finally redissolved in 0.4 mol l^{-1} HNO_3 + 0.04 mol l^{-1} HF for U, Th and Hf, and 0.4 mol l^{-1} HNO_3 for Lu-Yb and were analysed on a Neptune MC-ICPMS at the Origins Laboratory of the University of Chicago.

About 90% of the U–Th–Hf solution was used for concentration and high-precision isotope analysis of Hf and the remaining 10% was used for the analysis of U and Th concentrations by IDMS. Samples were introduced to the Neptune instrument through an Apex-Q + Spiro TDM desolvation inlet system using a 100 µl min^{-1} PFA self-aspirating nebulizer. Uranium and Th concentration measurements were made in dynamic mode, with ^{236}U and ^{229}Th ion beams on the secondary electron multiplier (SEM) and ^{235}U, ^{238}U and ^{232}Th on Faraday collectors. Yield calibration between the SEM and Faraday cups was performed using the NIST4321c U standard solution (courtesy of F. Marcantonio, Texas A&M University), which has a $^{238}U/^{235}U$ ratio of 137.90. This ratio was also used for the calculation of mass bias by standard-sample-standard bracketing technique. Tail contributions (abundance sensitivity) from ^{238}U and ^{232}Th on lower-abundance isotopes were negligible. Procedural blanks were treated similarly to the samples and their contributions were determined by IDMS. Blank contributions for U and Th were measured at 10 pg and 13 pg, respectively. Hafnium and Lu blanks were 13 pg and 7 pg, respectively. The MC-ICPMS acquisition method for U, Th and Lu concentrations consisted of 1 block of 5 cycles of 4.2 s integration time. Each sample was analysed up to three times. Hafnium isotopic composition and concentration were established using a method that consisted of 1 block of 15 cycles of 8.4 s integration time to achieve higher precisions. The average $^{176}Hf/^{177}Hf$ ratio for JMC-475 ($n = 150$, courtesy of J. Patchett, University of Arizona) was 0.282159 ± 3 ($2\sigma/\sqrt{n}$). All $^{176}Hf/^{177}Hf$ ratios measured in meteorites from Table 1 were normalized to the conventionally accepted value of 0.282160 for JMC-475 (ref. 36). Although isobaric interferences on ^{180}Hf from ^{180}Ta and ^{180}W were minimal (<0.1%), these corrections were implemented. The details of concentration calculations and offline mass bias and interference corrections for Hf isotopes are presented in ref. 34. Because Lu has only two naturally occurring isotopes, Yb isotopes are commonly used to determine instrumental mass bias for Lu concentration measurements by isotope dilution[13,37–39]. Lutetium concentrations were calculated using the natural $^{173}Yb/^{171}Yb$ ratio of 1.132685 (ref. 37) for mass bias calculations. Uncertainties (95% confidence intervals) were calculated on the basis of seven replicate analyses of the Allende CV3 carbonaceous chondrite.

34. Pourmand, A. & Dauphas, N. Distribution coefficients of 60 elements on TODGA resin: application to Ca, Lu, Hf, U and Th isotope geochemistry. *Talanta* **81,** 741–753 (2010).
35. Horwitz, E. P., McAlister, D. R., Bond, A. H. & Barrans, R. E. Novel extraction of chromatographic resins based on tetraalkyldiglycolamides: characterization and potential applications. *Solvent Extract. Ion Exch.* **23,** 319–344 (2005).
36. Vervoort, J. D. & Blichert-Toft, J. Evolution of the depleted mantle: Hf isotope evidence from juvenile rocks through time. *Geochim. Cosmochim. Acta* **63,** 533–556 (1999).
37. Chu, N. C. *et al.* Hf isotope ratio analysis using multi-collector inductively coupled plasma mass spectrometry: an evaluation of isobaric interference corrections. *J. Anal. At. Spectrom.* **17,** 1567–1574 (2002).
38. Connelly, J. N., Ulfbeck, D. G., Thrane, K., Bizzarro, M. & Housh, T. A method for purifying Lu and Hf for analyses by MC-ICP-MS using TODGA resin. *Chem. Geol.* **233,** 126–136 (2006).
39. Vervoort, J. D., Patchett, P. J., Söderlund, U. & Baker, M. Isotopic composition of Yb and the determination of Lu concentrations and Lu/Hf ratios by isotope dilution using MC-ICPMS. *Geochem. Geophys. Geosyst.* **5,** Q11002, doi:10.1029/2004GC00072 (2004).

LETTER

doi:10.1038/nature10104

Improved measurement of the shape of the electron

J. J. Hudson[1], D. M. Kara[1], I. J. Smallman[1], B. E. Sauer[1], M. R. Tarbutt[1] & E. A. Hinds[1]

The electron is predicted to be slightly aspheric[1], with a distortion characterized by the electric dipole moment (EDM), d_e. No experiment has ever detected this deviation. The standard model of particle physics predicts that d_e is far too small to detect[2], being some eleven orders of magnitude smaller than the current experimental sensitivity. However, many extensions to the standard model naturally predict much larger values of d_e that should be detectable[3]. This makes the search for the electron EDM a powerful way to search for new physics and constrain the possible extensions. In particular, the popular idea that new supersymmetric particles may exist at masses of a few hundred GeV/c^2 (where c is the speed of light) is difficult to reconcile with the absence of an electron EDM at the present limit of sensitivity[2,4]. The size of the EDM is also intimately related to the question of why the Universe has so little antimatter. If the reason is that some undiscovered particle interaction[5] breaks the symmetry between matter and antimatter, this should result in a measurable EDM in most models of particle physics[2]. Here we use cold polar molecules to measure the electron EDM at the highest level of precision reported so far, providing a constraint on any possible new interactions. We obtain $d_e = (-2.4 \pm 5.7_{stat} \pm 1.5_{syst}) \times 10^{-28} e$ cm, where e is the charge on the electron, which sets a new upper limit of $|d_e| < 10.5 \times 10^{-28} e$ cm with 90 per cent confidence. This result, consistent with zero, indicates that the electron is spherical at this improved level of precision. Our measurement of atto-electronvolt energy shifts in a molecule probes new physics at the tera-electronvolt energy scale[2].

Just as a magnetic dipole moment $\boldsymbol{\mu}$ in a magnetic field $\mathbf{B}$ has an energy $-\boldsymbol{\mu} \cdot \mathbf{B}$, an electric dipole moment $\mathbf{d}$ in an electric field $\mathbf{E}$ has an energy $-\mathbf{d} \cdot \mathbf{E}$ in the non-relativistic limit. A permanent EDM of the electron must lie along its spin[6], $\boldsymbol{\sigma}$, that is, $\mathbf{d} = d_e\boldsymbol{\sigma}$, making the electron's energy depend on whether the spin is parallel or antiparallel to $\mathbf{E}$. In an atom or molecule with an unpaired valence electron, the interaction of the electron EDM with an applied electric field results in an energy difference between two states that differ only in their spin orientation. This energy difference is proportional to d_e and changes sign when the direction of the field is reversed. A sensitive method of measuring this energy difference is to align the spin perpendicular to the field and measure its precession rate, which is proportional to the energy difference. An alternative description of the method is in terms of an interferometer. There is quantum interference between the two spin states, and the EDM appears as an interferometer phase shift that changes sign when the electric field is reversed.

To improve on the previous limit[7] we developed a technique using the dipolar molecule YbF (ref. 8) instead of the spherical Tl atom. This has two great advantages. First, at our modest operating field the interaction energy[9–15] of YbF due to d_e is 220 times larger than that obtained using Tl in a much larger field[7]. Second, the motional magnetic field, a source of systematic error that plagued the Tl experiment, has a negligible effect on YbF (ref. 8). Because of these advantages, it is possible to improve on the Tl experiment by using YbF molecules, even though the molecules are produced in much smaller numbers. A number of other EDM measurements, based on electron spin precession in atoms, molecules, molecular ions or solids, are in progress[4].

Figure 1 shows the interferometer apparatus[16]. Pulses of YbF molecules are emitted by the source[17]. The experiment uses those molecules in the $F = 0$ and $F = 1$ hyperfine levels of the ground state. The molecules pass through a first fluorescence detector, the pump detector, which simultaneously measures and empties out the $F = 1$ population. Then they enter a pair of electric field plates, between which are static electric and magnetic fields $(E, B)\hat{\mathbf{z}}$, where $\hat{\mathbf{z}}$ is the unit vector in the z direction (Fig. 1). This region is magnetically shielded. A radio-frequency (r.f.) pulse is applied to transfer molecules from $|F, m_F\rangle = |0, 0\rangle$ to the state $\frac{1}{\sqrt{2}}(|1, +1\rangle + |1, -1\rangle)$, where m_F is the component of the total angular momentum, F, along the z-axis. The molecules then evolve freely for a time T, during which the $m_F = \pm 1$ components develop a phase difference of $2\phi = 2(\mu_B B - d_e E_{eff})T/\hbar$, where μ_B is the Bohr magneton. This is due to the Zeeman shift $+\mu_B B m_F$ (ref. 18) and to the EDM shift expressed by the effective interaction $-d_e E_{eff} m_F$ (see Methods). A second r.f. pulse is then applied, resulting in a final $F = 0$ population proportional to $\cos^2\phi$, which the second fluorescence detector subsequently measures. For every pulse of molecules, the time-resolved signals from the pump and probe detectors are recorded; an example probe signal is shown in Fig. 2.

Scanning the phase difference via the magnetic field generates an interference curve, shown in Fig. 3. Reversal of the applied electric field produces a small phase shift $\delta\phi = 2d_e E_{eff} T/\hbar$, leading to a change in the detector count of $\delta I = (\mathrm{d}I/\mathrm{d}\phi)\delta\phi$. This is maximized by operating the interferometer at $B = \pm 13.6$ nT, which corresponds to $\phi = \pm\pi/4$, the steepest points on either side of the central fringe (Fig. 3). The intensity change is opposite on the two sides of the fringe because the slopes are opposite. Thus the EDM signal δI is the part of the fluorescence count that is correlated with the sign of $E \cdot B$. We calibrate the slope $\mathrm{d}I/\mathrm{d}\phi$ by making a step $\delta B = \pm 1.7$ nT in magnetic-field magnitude, and this too is done on each side of the central fringe. In addition to E, B and δB, several other parameters are switched in the experiment. The laser frequency is stepped by ± 340 kHz, the frequencies of the two r.f. pulses (ν_{rf1} and ν_{rf2}) are independently stepped by ± 1.5 kHz, their amplitudes (a_{rf1} and a_{rf2}) are independently stepped by $\pm 5\%$, and the phase difference (Φ_{rf}) between them is stepped around a randomly chosen value, ϕ_0, by $\pm\pi/2$. A computer places the machine in a new switch state before every beam pulse. The measurements are grouped into 'blocks' of 4,096 beam pulses, over which all 512 combinations of switch states are covered equally. Error signals, derived from each

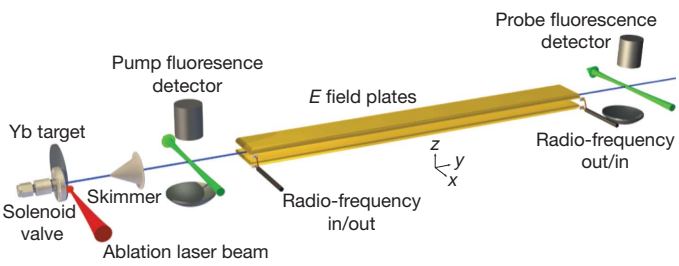


Figure 1 | Schematic diagram of the pulsed molecular beam apparatus.

[1]Centre for Cold Matter, Blackett Laboratory, Imperial College London, Prince Consort Road, London SW7 2AZ, UK.

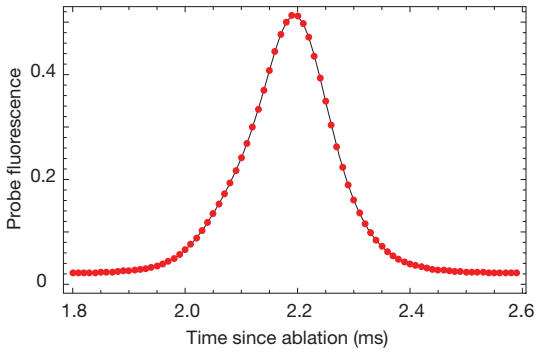


Figure 2 | Fluorescence from a typical beam pulse, measured on the probe detector.

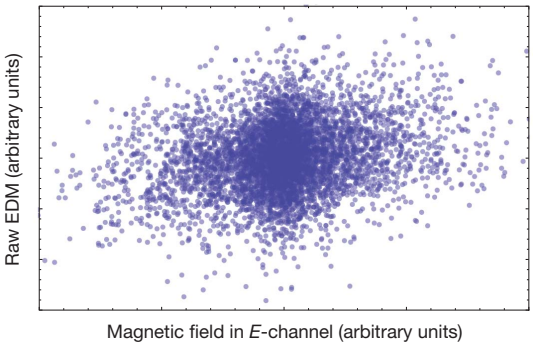


Figure 4 | The magnetic field correlated with the E reversal, measured at the fluxgate magnetometer, versus the EDM values. A slope is evident. The majority of measurements are not significantly perturbed by the magnetic field, but a small fraction do benefit from correction.

block of data, are fed back to the switched parameters to keep them switching around their optima.

Our measurement is derived from 6,194 blocks of data taken in 2010, comprising 25 million molecular beam pulses, together with many subsidiary measurements used to search for systematic errors. To analyse the data, we select the central 130 μs of each probe pulse (Fig. 2) and normalize it pulse by pulse to the pump fluorescence. This minimizes the effect of fluctuations of the molecular beam intensity. We calculate how much of the gated, normalized fluorescence signal is correlated with all 512 possible combinations of the modulated parameters. These correlations are called 'channels' and are denoted by $\{X\}$, where X indicates the parameter (or parameter combination) being modulated. The EDM phase shift, normalized to the shift from the small magnetic field step δB, is $\{E \cdot B\}/\{\delta B\}$. The other channels are valuable in elucidating the operation of the apparatus. Throughout the investigation the EDM values were concealed by adding a fixed unknown offset, which was only removed once the data collection and analysis were complete.

The EDM values obtained from the set of blocks are almost normally distributed but there tend to be a few more points in the wings of the distribution than in a normal distribution. The same is true of other quantities of interest that we extract from the data. For all these quantities, we calculate the 5% trimmed mean[19], a simple robust statistic that drops the largest and smallest 5% of the data. We use the bootstrap method[20] to determine the associated statistical uncertainty. For non-normal distributions, these methods give more reliable measures than the mean and standard error.

Fluctuations in the ambient magnetic field of the laboratory inevitably have some component that is, by chance, synchronous with the switching pattern of E. This contributes a little to the noise in the EDM, as shown in Fig. 4, though not to the long-time average value. We suppress this excess noise by correcting the EDM, block by block, according to the magnetic field readings of a magnetometer (Methods). The central value and statistical uncertainty of this magnetic field correction are given in Table 1. The correction has a negligible effect on the central value of the EDM but reduces the statistical error by 3.5%.

We find that the phase of the interferometer shifts linearly with the detunings of the r.f. pulses at a rate of $(283 \pm 6) \times 10^{-9}$ rad Hz^{-1} for the first r.f. pulse, and $(-94 \pm 5) \times 10^{-9}$ rad Hz^{-1} for the second r.f. pulse. If the magnitude of the electric field changes when E is reversed, then through the Stark shift, the r.f. transition frequency changes. This results in a change in the interferometer phase that correlates with E, mimicking the EDM phase. This systematic error can be corrected using the information contained in every block of data. The phase change resulting from a detuning of the first r.f. pulse is measured by $\{v_{rf1} \cdot B\}$, and the change in the detuning resulting from the change in electric field magnitude is measured by $\{v_{rf1} \cdot E\}$. The product of these two channels, together with a calibration factor that we have measured, determines the EDM-like phase due to the E-correlated detuning of the first r.f. transition, and we use this to apply a correction to each block of data. A similar correction is made for the second r.f. pulse. The central values and statistical uncertainties of the two r.f. phase corrections are given in Table 1. As an additional check, we made measurements in which we deliberately change the r.f. frequency when we switch E. We see that the resulting systematic error is entirely removed once the corrections are applied to these data, thus verifying the correction procedure.

There are several sources of systematic uncertainty on the EDM measurement that must be considered. First, there may be systematic effects, other than the r.f.-induced phases described above, caused by a change in field magnitude when E reverses. We investigate this by changing the field magnitude intentionally by δE when the field switches. Once the r.f. phase corrections are applied to these data,

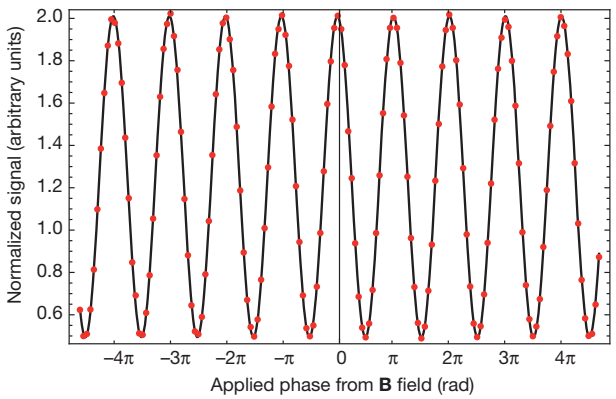


Figure 3 | Interferometer fringes produced by magnetic field scan. Dots indicate the probe fluorescence normalized to the pump fluorescence. The line is the fit to the cosine-squared model.

Table 1 | Summary of applied corrections and uncorrected systematic uncertainties

	Correction	Statistical	Systematic
Magnetic-field correction	−0.3	1.7	<0.1
rf1 phase correction	5.0	0.9	<0.1
rf2 phase correction	0.5	0.7	<0.01
Uncorrected δE effects	–	–	1.1
$\bar{V}$ uncertainty	–	–	0.1
$\{v_{rf1}\}$ correlation	–	–	1.0
Geometric phase	–	–	0.03
Leakage currents	–	–	0.2
Shield magnetization	–	–	0.25
$\mathbf{v} \times \mathbf{E}$ effect	–	–	0.0005

The units are $10^{-28} e$ cm. The statistical uncertainty on the corrections gives a measure of their random spread over the whole data set. In the final analysis the corrections are applied block-by-block, so these statistical uncertainties are naturally incorporated in the final EDM statistical uncertainty. The systematic uncertainty in the corrections is negligible.

we find no evidence of any residual systematic EDM that depends on δE. The upper bound on the gradient of any such systematic, with respect to δE, is $-11 \times 10^{-28} e\,\mathrm{cm/(V\,cm^{-1})}$. In the r.f. regions we measure asymmetries δE of approximately 100 mV cm^{-1} and we take this to be typical throughout the interaction region. Combining this level of asymmetry with the worst-case slope above gives a systematic uncertainty of $1.1 \times 10^{-28} e$ cm (Table 1).

Electric-field-plate potentials that are not symmetric around the ground potential are another possible source of systematic error. We characterize this in terms of the mean potential $\bar{V}$ of the two electric field plates relative to the surrounding grounded apparatus. Near the edges of the plates, the field does not point entirely along $\hat{\mathbf{z}}$, but the direction of the field reverses perfectly as long as $\bar{V} = 0$. However, when $\bar{V} \neq 0$ the reversal is imperfect, and this, coupled with other imperfections, may result in a systematic error. We investigate this by deliberately applying large mean potentials of $\bar{V} = -1{,}000.5$ V and $\bar{V} = +1{,}015.0$ V, and we find from this data a systematic shift with a slope of $(0.099 \pm 0.016) \times 10^{-28} e\,\mathrm{cm\,V^{-1}}$. The plate potentials used for our data set are measured to have a mean voltage of less than 1 V. This results in a systematic uncertainty of $0.1 \times 10^{-28} e$ cm.

A study of the data taken at non-zero $\bar{V}$ revealed an unexplained correlation between the measured EDM and the frequency detuning of the first r.f. pulse. Unlike the effect described above, this systematic effect does not depend on δE. We see no evidence of the effect in the data taken at $\bar{V} = 0$. Nonetheless, by considering the worst-case correlation consistent with the $\bar{V} = 0$ data, and the measured average frequency detuning of the first r.f. pulse, we calculate a conservative systematic uncertainty of $1 \times 10^{-28} e$ cm.

The direction of the electric field in the rest frame of the molecules rotates slightly as they move through the apparatus. This induces a geometric interferometer phase that can result in a systematic error[21]. We calculate an upper limit on this effect (see Supplementary Information) of $3 \times 10^{-30} e$ cm.

Magnetic fields generated inside the magnetic shields that reverse with the electric field are a potential source of systematic error. These magnetic fields are not well sensed by the magnetometers, which are outside the inner layer of magnetic shielding. We consider the three mechanisms that could generate such fields:

(1) Leakage current to the high-voltage plates. The current flowing to or from each electric field plate is monitored[22] throughout the experiment. The component that reverses synchronously with E is less than 1 nA averaged over the EDM data set. A most conservative estimate (see Supplementary Information) of the possible false EDM given by these currents is $0.2 \times 10^{-28} e$ cm.

(2) Inner-shield magnetization. It is possible that the plate-charging currents could magnetize the shields, generating a magnetic field that reverses with E. We have determined this field by pulsing a hundred times the normal current through a similar shield set-up on the bench and measuring the resulting field with a fluxgate magnetometer. We deduce that the false EDM due to shield magnetization is $(-0.16 \pm 0.17) \times 10^{-28} e$ cm. As this is consistent with zero, we do not make any correction to the measured EDM, but allow a systematic uncertainty of $0.25 \times 10^{-28} e$ cm.

(3) Motional magnetic field. The laboratory-frame electric field has a magnetic component in the rest frame of the molecules $\mathbf{B}_\mathrm{m} = \mathbf{E} \times \mathbf{v}/c^2$, where $\mathbf{v}$ is the velocity of the molecules with respect to the apparatus. This can produce a false EDM if there is also a stray magnetic field B_y. This was a limiting systematic error in ref. 7. The effect is strongly suppressed in our case because of the large (8 MHz) tensor Stark splitting of the $F = 1$ manifold, which renders the molecule insensitive to magnetic fields in the x–y plane, as discussed in ref. 8. Our stray B_y is everywhere less than 30 nT, which gives a calculated false EDM of less than $5 \times 10^{-32} e$ cm. We have also checked empirically that the addition of a 500 nT transverse field produces no evident effect.

A number of other consistency checks and searches for systematic errors were made and are described in detail in the Supplementary Information.

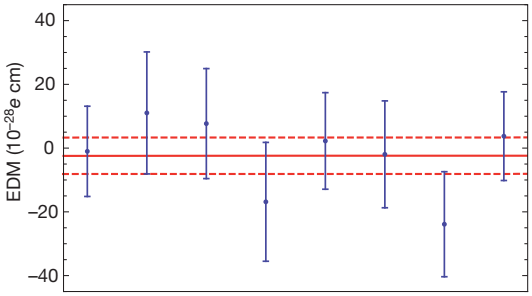


Figure 5 | EDM values for each manual-reversal state of the machine. The error bars indicate the 68% confidence level. The most important manual reversal is the electric-field reversal: the first four points correspond to one electric-field configuration, and the last four to the other. The solid and dashed lines show the mean value and its statistical error.

In addition to the computer-controlled switches, we make three manual reversals. The high-voltage connections are swapped to reverse **E**, the magnet wires are interchanged to reverse **B** and the r.f. cables are swapped to reverse the direction of r.f. propagation along the field plates. These manual changes are made infrequently—typically one switch per day—and they are valuable in identifying and eliminating systematic effects. Roughly equal numbers of blocks are taken in all eight of the manual states. When we divide the data according to these manual-reversal states and analyse each data set separately, the EDMs obtained are consistent with one another, as shown in Fig. 5. We also divide the data according to the polarization angles of the pump and probe and find no correlation with either.

Combining the systematic uncertainties in quadrature yields the final result $d_\mathrm{e} = (-2.4 \pm 5.7_\mathrm{stat} \pm 1.5_\mathrm{syst}) \times 10^{-28} e$ cm, where the first uncertainty is statistical (68% symmetric confidence interval[23]) and the second systematic. This is consistent with zero and with the previous best measurement[7]. The result is 54 times more precise than our previous measurement[8]. Treating the statistical and systematic errors on equal terms, we can extract an upper bound on the size of the EDM of $|d_\mathrm{e}| < 10.5 \times 10^{-28} e$ cm with 90% confidence. This is 1.5 times smaller than the previous upper limit[7].

Our error is dominated by the statistical uncertainty of the measurement. The limiting systematic errors in the measurement are sufficiently well understood that we can readily reduce them to the $10^{-29} e$ cm range. Our experiment leads the way in the application of cold molecule techniques to precision measurement and we are well placed to take advantage of recent advances in the preparation[24–26] and control[27] of cold molecules to improve our measurement precision. This will allow us to probe for new particle physics at tens of tera-electronvolts.

METHODS SUMMARY

Pulses of YbF are emitted by the source[17] every 40 ms and travel through the magnetically shielded apparatus (Fig. 1) at a speed of 590 m s^{-1}. The pump detector depletes and detects the $F = 1$ population while the probe detector measures the $F = 0$ population. Two r.f. π-pulses, separated by the free-evolution time T, and tuned to the Stark-shifted hyperfine interval near 170 MHz, coherently transfer molecules between the $F = 0$ and $F = 1$ states. The primary signal is the detected $F = 0$ population, which is proportional to $\cos^2\phi$. The electron EDM is obtained from the part of ϕ that correlates with the sign of E, which in turn is obtained from the signal correlating with the sign of $E \cdot B$.

To measure this correlation, and a rich set of other signal correlations, the machine is put into a new state between each beam pulse. There are nine switched parameters, and hence 512 different switch combinations; each is set eight times in every data block (a group of 4,096 pulses). For each block, the switching sequence is chosen at random from a set of possible sequences; all of these switch B frequently to eliminate magnetic field noise, switch E infrequently to minimize the dead time associated with this switch, and switch $E \cdot B$ aperiodically to eliminate signal drifts from this channel[28]. Between one block and the next, the relative phase of the two r.f. pulses is randomly changed, the linear polarizations of pump and probe are randomly rotated, and the central values of the magnetic field, the laser frequency, and the frequencies and amplitudes of the two r.f. pulses, are adjusted towards their ideal values.

Diagnostic data are obtained from a fluxgate magnetometer placed between the two shields, three other magnetometers around the laboratory, and two ammeters[22] that measure the currents flowing to the electric field plates.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 18 February; accepted 8 April 2011.

1. Khriplovich, I. B. & Lamoreaux, S. K. *CP Violation Without Strangeness* (Springer, New York, 1997).
2. Pospelov, M. & Ritz, A. Electric dipole moments as probes of new physics. *Ann. Phys.* **318,** 119–169 (2005).
3. Commins, E. D. Electric dipole moments of leptons. In *Advances in Atomic, Molecular, and Optical Physics* Vol. 40, 1–56 (eds Bederson, B. & Walther, H.), Academic Press (1999).
4. Commins, E. D. & DeMille, D. in *Lepton Dipole Moments* (eds Roberts, B. L. & Marciano, W. J.) Ch. 14 (World Scientific, Singapore, 2010).
5. Sakharov, A. D. Violation of CP invariance, C asymmetry, and baryon asymmetry of the Universe. *Pis'ma ZhETF* **5,** 32–35 (1967); *Sov. Phys. JETP Lett.* **5,** 24–27 (1967).
6. Edmonds, A. R. *Angular Momentum in Quantum Mechanics* 73–77 (Princeton University Press, 1996).
7. Regan, B. C., Commins, E. D., Schmidt, C. J. & DeMille, D. New limit on the electron electric dipole moment. *Phys. Rev. Lett.* **88,** 071805 (2002).
8. Hudson, J. J., Sauer, B. E., Tarbutt, M. R. & Hinds, E. A. Measurement of the electron electric dipole moment using YbF molecules. *Phys. Rev. Lett.* **89,** 023003 (2002).
9. Hinds, E. A. Testing time reversal symmetry using molecules. *Phys. Scr.* **T70,** 34–41 (1997).
10. Kozlov, M. G. & Ezhov, V. F. Enhancement of the electric dipole moment of the electron in the YbF molecule. *Phys. Rev. A* **49,** 4502–4507 (1994).
11. Kozlov, M. G. Enhancement of the electric dipole moment of the electron in the YbF molecule. *J. Phys. B* **30,** L607–L612 (1997).
12. Titov, A., Mosyagin, M. & Ezhov, V. P,T-odd spin-rotational Hamiltonian for YbF molecule. *Phys. Rev. Lett.* **77,** 5346–5349 (1996).
13. Quiney, H. M., Skaane, H. & Grant, I. P. *Hyperfine and PT-odd effects in YbF* $^2\Sigma$. *J. Phys. B* **31,** L85–L95 (1998).
14. Parpia, F. A. Ab initio calculation of the enhancement of the electric dipole moment of an electron in the YbF molecule. *J. Phys. B* **31,** 1409–1430 (1998).
15. Mosyagin, N., Kozlov, M. & Titov, A. Electric dipole moment of the electron in the YbF molecule. *J. Phys. B* **31,** L763–L767 (1998).
16. Hudson, J. J. *et al.* Pulsed beams as field probes for precision measurement. *Phys. Rev. A* **76,** 033410 (2007).
17. Tarbutt, M. R. *et al.* A jet beam source of cold YbF radicals. *J. Phys. B* **35,** 5013–5022 (2002).
18. Ma, T., Butler, C., Brown, J. M., Linton, C. & Steimle, T. C. Optical Zeeman spectroscopy of ytterbium monofluoride, YbF. *J. Phys. Chem. A* **113,** 8038–8044 (2009).
19. Maronna, R. A., Martin, D. R. & Yohai, V. J. *Robust Statistics; Theory and Methods* 31–32 (Wiley, 2006).
20. Efron, B. & Tibshirani, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* **1,** 54–75 (1986).
21. Tarbutt, M. R., Hudson, J. J., Sauer, B. E. & Hinds, E. A. Prospects for measuring the electric dipole moment of the electron using electrically trapped polar molecules. *Faraday Discuss.* **142,** 37–56 (2009).
22. Sauer, B. E., Kara, D. M., Hudson, J. J., Tarbutt, M. R. & Hinds, E. A. A robust floating nanoammeter. *Rev. Sci. Instrum.* **79,** 126102 (2008).
23. Efron, B. Better bootstrap confidence intervals. *J. Am. Stat. Assoc.* **82,** 171–185 (1987).
24. Skoff, S. M. *et al.* Diffusion, thermalization and optical pumping of YbF molecules in a cold buffer gas cell. *Phys. Rev. A* **83,** 023418 (2011).
25. Barry, J. F., Shuman, E. S. & DeMille, D. A bright, slow cryogenic molecular beam source for free radicals. Preprint at ⟨http://arxiv.org/abs/1101.4229⟩ (2011).
26. Hutzler, N. R. *et al.* A cryogenic beam of refractory, chemically reactive molecules with expansion cooling. Preprint at ⟨http://arxiv.org/abs/1101.4217⟩ (2011).
27. van de Meerakker, S. Y. T., Bethlem, H. L. & Meijer, G. in *Cold Molecules: Theory, Experiment, Applications* (eds Krems, R., Stwalley, W. & Friedrich B.) Ch. 14 (CRC Press, 2009).
28. Harrison, G. E., Player, M. A. & Sandars, P. G. H. A multichannel phase-sensitive detection method using orthogonal square waveforms. *J. Phys. E* **4,** 750–754 (1971).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge the contributions of P. Condylis and H. Ashworth. We are grateful for technical assistance from J. Dyne and V. Gerulis. This work was supported by the UK research councils STFC and EPSRC, and by the Royal Society. J.J.H. is supported by an STFC Advanced Fellowship.

Author Contributions J.J.H. was involved in all aspects of the measurement, led the analysis, and drafted the manuscript. D.M.K. developed many of the systematic tests, worked on taking the data set, and contributed to the analysis. I.J.S. had primary responsibility for taking the data set, and contributed to the development of the data acquisition techniques. B.E.S. was involved in all aspects of the measurement, and designed much of the hardware. M.R.T. built the molecular beam source, contributed to the analysis, and drafted the manuscript. E.A.H. contributed to the analysis, drafted the manuscript and led the team. All authors discussed the results, improved the manuscript and were equally involved in setting the direction of the work.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to E.A.H. (ed.hinds@imperial.ac.uk).

METHODS

Apparatus. The apparatus is shown in Fig. 1. A solenoid valve opens every 40 ms to release a pulse of Ar containing 2% SF_6. Ytterbium atoms laser-ablated from a target beside the valve react with the gas pulse to form YbF. The gas expands, cools and is skimmed to form a beam with a temperature of 3 K and a centre-of-mass velocity of 590 m s^{-1} (ref. 17). The YbF molecules are mainly in the electronic and vibrational ground state $X^2\Sigma^+(v=0)$. Those in the rotational ground state are distributed over the hyperfine levels $F=0$ and $F=1$, separated by 170.254 MHz. A single-mode continuous-wave dye laser provides the linearly polarized pump and probe beams shown in Fig. 1. The pump and probe are respectively tuned to the $F=1$ and $F=0$ components of the A–X $Q(0)$ transition, so that the pump empties out the $F=1$ population and the probe measures the $F=0$ population by laser-induced fluorescence detection. Each packet of molecules passing through the probe beam generates a current pulse in the photomultiplier corresponding to ~5,000 detected photons. The current pulse is digitized in 80 bins over 800 μs to produce signals such as that shown in Fig. 2. The pump fluorescence is recorded in a similar way. We also record the intensities of both laser beams. The timing of the experiment is phase-locked to the mains electrical supply.

The field plates are gold-coated cast aluminium, 75 cm long, 7 cm wide and 1.2 cm apart. The static electric and magnetic fields between these plates are typically $E=\pm 10$ kV cm^{-1} and $B=\pm 13$ nT. The plate structure doubles as a TEM (transverse electromagnetic) transmission line to propagate 170 MHz radiation in either direction. The r.f. pulses are designed to be π-pulses, so that the transfer of molecules between the $|F,m_F\rangle=|0,0\rangle$ state and the $\frac{1}{\sqrt{2}}(|1,+1\rangle+|1,-1\rangle)$ state occurs with unit efficiency. The first r.f. pulse is applied 1.1 ms after the ablation pulse, when the molecules are approximately 13 cm inside the plates. The second r.f. pulse is applied after the free evolution time of $T=642$ μs. Both pulses are 18-μs-long r.f. magnetic field pulses polarized along $\hat{\mathbf{x}}$ (Fig. 1). If the π-pulses are imperfect, coherence between $F=0$ and $F=1$ states results in additional, unwanted interference terms. We suppress these terms by averaging the relative r.f. phase $\phi_0 \pm \pi/2$ over the Φ_{rf} switch and by randomizing ϕ_0 between blocks. The theory of two-pulse r.f. transitions within this three-level manifold is developed fully in section IV.B of ref. 29.

The beam line is enclosed by two layers of magnetic shielding. The high-voltage feeds pass close together through a single hole in the inner magnetic shield near the centre of the plates to minimize shield magnetization by the charging currents. A fluxgate magnetometer between the shields measures the magnetic field parallel to $\hat{\mathbf{z}}$ near the probe detector. Three other magnetometers of lower sensitivity are used to monitor the laboratory magnetic field—one near the beam machine, one close to the high-voltage relays that reverse $\mathbf{E}$, and one close to the computer interface that controls the experiment. These are also read after every pulse and their primary purpose is to ensure that E-reversal does not generate a magnetic field. The same analogue–digital converter board that reads these signals also monitors two dummy voltages, a battery and a short circuit. These are used to check that there are no systematic errors in the signal processing electronics and data analysis.

Diagnostic data are also obtained from two ammeters[22] that measure the currents flowing to the electric field plates.

Characterizing the machine. We have mapped the spatial variation of the electric, magnetic and r.f. fields, as described in ref. 16. We find that the electric field varies by roughly 1% over the length of the plates, and that the ambient magnetic field is typically less than 10 nT throughout the region that we use for the interferometer. The r.f. field has a small standing-wave ratio, corresponding to a 4% power reflection coefficient at each end. In the TEM mode, the r.f. electric field is constrained by the same boundary conditions as the static field, ensuring that the r.f. magnetic field is accurately perpendicular to E and to the propagation direction. The r.f. field at each end of the plates has some ellipticity, due to the transient where the transmission line is coupled to coaxial cable. This decays away over a few centimetres.

Switching sequence. As discussed in the main text, nine separate parameters are switched in the experiment. A set of 4,096 beam pulses forms a block of data, within which all 512 combinations of switch states are covered equally. The sequence of switches applied within a block, known as the switching pattern, must satisfy three requirements. First, the magnetic field should switch frequently to eliminate magnetic field noise. Second, the electric field must switch less often because E reversal incurs a dead time of 14 s. This allows time to discharge and recharge the plates while keeping the transient currents below 5 μA to avoid magnetizing the shields. By the time we restart data acquisition the current is close to its steady value of ~1 nA. This restriction is important because a magnetic field reversing with E can generate a systematic error. Third, the switching sequence of $E \cdot B$ should be as aperiodic as possible so that signal drifts do not influence this channel[28]. Within these restrictions, there are still a large number of possible switching patterns from which the computer randomly chooses one at the start of every block. At the end of each block the channel values are calculated and some of these are used to optimize the running of the machine. For example, $\{B\}$ measures how well the operating fields are centred around $B=0$ and this provides an error signal at the end of each block that is fed back to compensate for small drifts of the ambient field. Similarly, $\{v_{rf1}\}$ and $\{v_{rf2}\}$ are used to lock the r.f. frequencies to resonance while $\{a_{rf1}\}$ and $\{a_{rf2}\}$ are used to lock the r.f. amplitudes to the π-pulse condition. The laser- frequency channel {LF} is used to keep the laser on resonance. Between blocks the mean relative phase ϕ_0 between the two r.f. pulses is randomly changed and the linear polarizations of the pump and probe laser beams are randomly rotated. Including the dead time, each block takes approximately 6 min to accumulate.

EDM interaction. The interaction of the electron in a molecule with an applied electric field is more complicated than that of a free electron, described in the introduction. It is possible however to write the interaction as $-\mathbf{d}\cdot\mathbf{E}_{eff}$. The effective electric field $\mathbf{E}_{eff}$, which depends nonlinearly on the applied electric field, accounts for the complexity of the molecular environment. Under our operating conditions the effective field has magnitude 14.5 GV cm^{-1} and is aligned antiparallel to the applied field[10–15]. Thus, the energy shift of the $(F=1,m_F)$ state of the molecule due to the electron EDM is $-d_e E_{eff} m_F$ where $E_{eff}=-14.5$ GV cm^{-1}. In deriving the EDM we have assumed that the effective field is known exactly. Although there is some uncertainty in the theoretical calculation, even an uncertainty of 10% would have no impact on our error at the level reported here.

29. Tarbutt, M. R., Hudson, J. J., Sauer, B. E. & Hinds, E. A. in *Cold Molecules: Theory, Experiment, Applications* (eds Krems, R., Stwalley, W. & Friedrich B.) Ch. 15 (CRC Press, 2009).

LETTER

doi:10.1038/nature10013

Interannual atmospheric variability forced by the deep equatorial Atlantic Ocean

Peter Brandt[1], Andreas Funk[1], Verena Hormann[1]†, Marcus Dengler[1], Richard J. Greatbatch[1] & John M. Toole[2]

Climate variability in the tropical Atlantic Ocean is determined by large-scale ocean–atmosphere interactions, which particularly affect deep atmospheric convection over the ocean and surrounding continents[1]. Apart from influences from the Pacific El Niño/Southern Oscillation[2] and the North Atlantic Oscillation[3], the tropical Atlantic variability is thought to be dominated by two distinct ocean–atmosphere coupled modes of variability that are characterized by meridional[4,5] and zonal[6,7] sea-surface-temperature gradients and are mainly active on decadal and interannual timescales, respectively[8,9]. Here we report evidence that the intrinsic ocean dynamics of the deep equatorial Atlantic can also affect sea surface temperature, wind and rainfall in the tropical Atlantic region and constitutes a 4.5-yr climate cycle. Specifically, vertically alternating deep zonal jets of short vertical wavelength with a period of about 4.5 yr and amplitudes of more than 10 cm s^{-1} are observed, in the deep Atlantic, to propagate their energy upwards, towards the surface[10,11]. They are linked, at the sea surface, to equatorial zonal current anomalies and eastern Atlantic temperature anomalies that have amplitudes of about 6 cm s^{-1} and 0.4 °C, respectively, and are associated with distinct wind and rainfall patterns. Although deep jets are also observed in the Pacific[12] and Indian[13] oceans, only the Atlantic deep jets seem to oscillate on interannual timescales. Our knowledge of the persistence and regularity of these jets is limited by the availability of high-quality data. Despite this caveat, the oscillatory behaviour can still be used to improve predictions of sea surface temperature in the tropical Atlantic. Deep-jet generation and upward energy transmission through the Equatorial Undercurrent warrant further theoretical study.

Tropical Atlantic variability, which modulates the seasonal migration of the intertropical convergence zone, is dominated by two modes of behaviour[8,9]. The meridional mode, peaking during boreal spring, is characterized by a north–south sea-surface-temperature (SST) gradient that drives cross-equatorial wind anomalies from the cold hemisphere to the warm[4,5]. The zonal mode is characterized by an east–west SST gradient along the Equator and is associated with marked zonal wind anomalies[6,7]. It is most pronounced during boreal summer when the seasonal maximum in equatorial upwelling leads to the development of the eastern Atlantic SST cold tongue. The zonal mode is often referred to as the Atlantic counterpart to the Pacific El Niño. The period of zonal-mode-like oscillations estimated from observations, models and theory ranges from 19 months to 4 years[6,14–16]. However, aspects of the intrinsic ocean dynamics, such as year-to-year variations in the strength of tropical instability waves, are similarly identified as causes of interannual SST variability[17] and may themselves be able to force variability in the atmosphere.

During the past 10–20 yr, the eastern equatorial Atlantic SST, represented by the ATL3 index (that is, the average SST anomaly inside the box shown in Fig. 1a), has shown pronounced variability on interannual timescales, dominated by the period range of 4–5 yr; maximum explained variance of different ocean parameters is found at a period of 1,670 d (Supplementary Fig. 1). The associated harmonic amplitude of local SST fluctuations, which is 0.29 ± 0.08 °C averaged over the ATL3 region, is generally high in the eastern equatorial Atlantic, with local amplitudes of up to 0.4 °C (Fig. 1a and Supplementary Fig. 2). The regression of surface winds and rainfall on the 1,670-d SST harmonic reveals that anomalous westerlies along the Equator, convergent meridional wind anomalies particularly in the western tropical Atlantic, and positive rainfall anomalies in a wide belt around the Equator are associated with positive SST anomalies.

A 1,670-d cycle is also found in the surface geostrophic zonal velocity anomaly at the Equator and is again the dominant interannual variability, with a harmonic amplitude of 5.9 ± 1.9 cm s^{-1}. Phases of eastward surface flow coincide with SST warm phases in the eastern equatorial Atlantic (Fig. 1b). Whereas the 1,670-d period stands out as the dominant interannual variability timescale of the equatorial zonal surface flow, this is not the case for the wind forcing, which instead shows more irregular fluctuations during the analysed time interval (NCEP/NCAR reanalysis wind data). Such a dominant signal in the ocean seems to contradict early model results, in which the equatorial ocean response to wind forcing with periods longer than about 150 d was found to be a succession of equilibrium responses with the strength of the flow independent of the forcing period[18]. As we show below, variability in the 4–5-yr period band is a ubiquitous feature of the equatorial Atlantic and, furthermore, is associated with upward propagation of energy in the ocean. We propose that the variability in the equatorial zonal surface flow is not due to wind forcing with the same period but rather is a mode internal to the ocean, with its origin in the abyss (perhaps as deep as several thousand metres). If this is indeed the case, then the observed atmospheric variability in the 4–5-yr period band in the equatorial Atlantic can be interpreted as a consequence of internal ocean dynamics.

Analysis of zonal velocities at 1,000-m depth as observed by Argo floats[19] reveals periodic behaviour similar to that of the SST and surface geostrophic zonal velocity anomalies (Fig. 1b). The dominant period, of 4.4 yr, in the Argo float drift data for the period 1998–2010 is in agreement with earlier estimates from moored zonal velocity observations in the depth range 600–1,800 m made during 2000–2006[11] (4.4 yr) and with the estimate from hydrographic observations made during 1972–1998[10] (5 ± 1 yr). The deep velocity and density fluctuations have been dynamically described as a mixture of high-baroclinic-mode Kelvin and Rossby waves representing quasi-steady equatorial deep jets[10,11]. Such vertically alternating zonal jets with vertical wavelengths between 300 and 700 m are similarly present in the Pacific[12,20] and Indian oceans[13,21]. In the Atlantic, a downward phase velocity of equatorial deep jets (of about 100 m yr^{-1}) is observed[11] that corresponds, according to linear internal wave theory, to upward energy propagation. Our moored observations reveal downward phase propagation from below the Equatorial Undercurrent (EUC) at about 200-m depth to about 2,000-m depth (Fig. 2 and Supplementary Fig. 3), suggesting a deep generation mechanism for equatorial deep jets. Observed variations in the vertical phase velocity are probably due to changes in the amplitudes of different superimposed baroclinic modes, as also indicated by changes in the vertical wavelength (Fig. 2). Theories of

[1]IFM-GEOMAR, Leibniz-Institut für Meereswissenschaften an der Universität Kiel, Düsternbrooker Weg 20, 24105 Kiel, Germany. [2]Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA. †Present address: Cooperative Institute for Marine and Atmospheric Studies, University of Miami, and National Oceanic and Atmospheric Administration/Atlantic Oceanographic and Meteorological Laboratory, Miami, Florida 33149, USA.

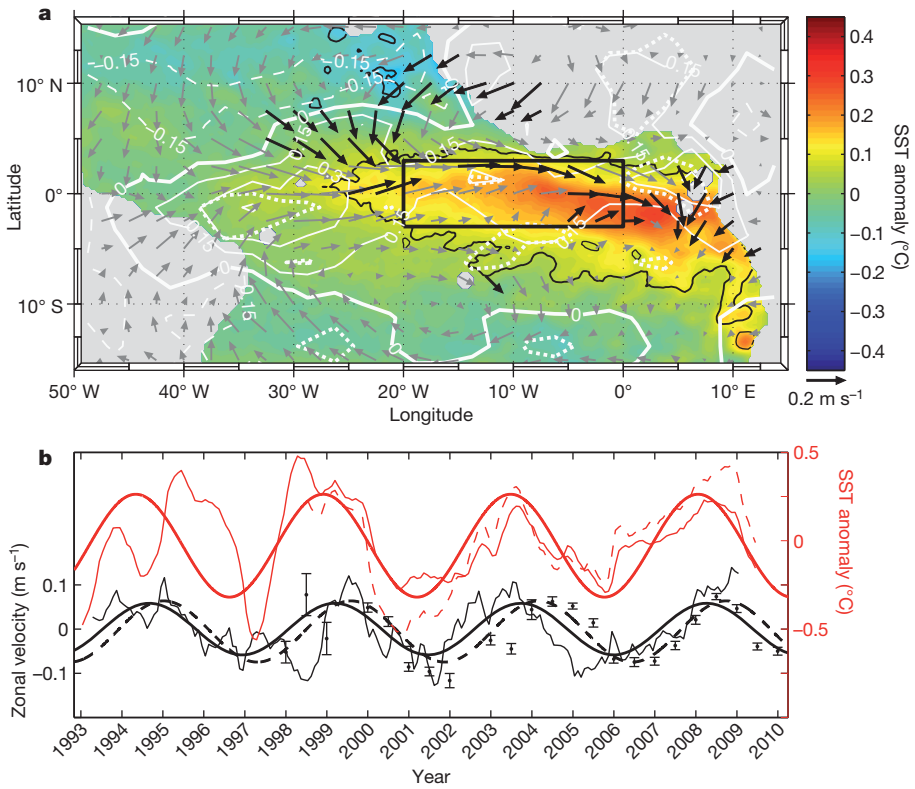


Figure 1 | Interannual variability in the tropical Atlantic associated with a 1,670-d cycle. **a**, Anomalies of SST (colour scale), surface wind (arrows) and rainfall (white contours: solid, positive; dashed, negative; every 0.15 mm d^{-1}) as determined through regression on the harmonic fit of the SST anomalies (microwave optimally interpolated SST) averaged within the marked box (ATL3: 3° S–3° N, 20° W–0°). We mark significant correlations (95%) of harmonic fit with SSTs (black lines), winds (black arrows) and rainfall (white dotted lines). **b**, ATL3 SST anomaly (microwave optimally interpolated SST, red dashed; HadISST, red thin solid) with 1,670-d harmonic fit (red thick solid), surface zonal velocity anomaly (Equator, 35° W–15° W; black thin solid) with 1,670-d harmonic fit (black thick solid), and 1,000-m zonal velocity (1° S–1° N, 35° W–15° W; black dots with standard errors) with 1,670-d harmonic fit (black thick dashed).

deep-jet generation involve instabilities associated with the propagation of intraseasonal mixed Rossby gravity waves[22,23] or the Equator-crossing deep western boundary current[24]. However, until now the proposed theories have failed to explain the observed strength and complex behaviour of the deep jets in the different ocean basins.

Propagation of deep-jet energy towards the surface is complicated by the presence of a strong, vertically-sheared mean current, the EUC, with maximum eastward velocities of more than 60 cm s^{-1} at about 80-m depth (Fig. 3a). Theoretical studies indicate that the EUC effectively modifies dispersion characteristics of Kelvin and Rossby waves[25]. On seasonal timescales, the background flow partly inhibits the downward propagation of high-baroclinic-mode energy, explaining the dominance of low-baroclinic-mode seasonal waves at depth. Theoretical studies of internal wave propagation motivated by observed internal wave transmissions across an atmospheric jet suggest, however, that an energy transfer across critical levels—that is, where the horizontal phase velocity equals the background mean flow—is possible[26].

The amplitude of the 1,670-d harmonic oscillation of zonal velocity in the upper 600 m of the water column is largest in the 300–600-m depth interval (Fig. 3b), where it explains up to 60% of the variance contained in the monthly zonal velocity anomalies (Fig. 3d). Local minima in the amplitude of the 1,670-d oscillation are indicated near

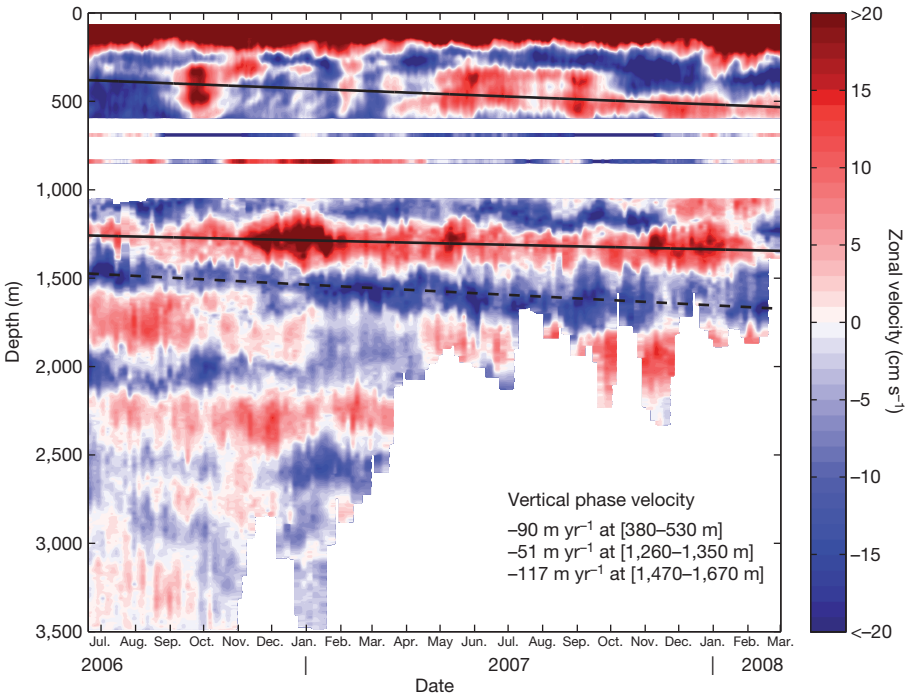


Figure 2 | Zonal velocities at the Equator, 23° W. Velocity data above 600 m are from a moored acoustic Doppler current profiler with annual and semi-annual cycles subtracted, those between 600 and 1,000 m are from two single-point current meters, and those below 1,000 m are from a moored profiler. The white areas mark depths not sampled by the deployed instrumentation. Linearized phase lines (eastward jets, solid; westward jet, dashed) of equatorial deep jets are calculated from about 7-yr of moored current data (above 600 m) and from the presented data (below 1,000 m). Associated vertical phase velocities are given in the figure.

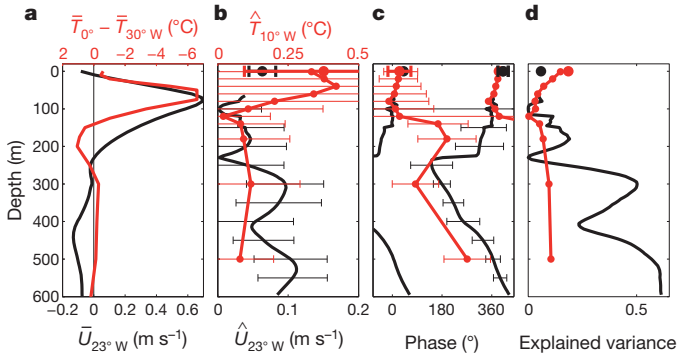


Figure 3 | Mean zonal velocity, zonal temperature gradient and harmonic analysis of 1,670-d oscillation. **a**, Moored mean zonal velocity ($\bar{U}$) at the Equator, 23 °W (black), and climatological[27] mean zonal temperature ($\bar{T}$) difference at the Equator between 0° and 30° W (red). **b**–**d**, 1,670-d harmonic amplitude (**b**), phase (**c**) and explained variance (**d**) of equatorial moored zonal velocities at 23° W (black curves ($\hat{U}_{23°\,W}$)), equatorial surface zonal velocity averaged between 35° W and 15° W (black dots), and subsurface temperatures (red curves ($\hat{T}_{10°\,W}$) and small red dots) and microwave optimally interpolated SST (big red dot) at the Equator, 10° W. Zero phase corresponds to 1 January 1993; explained variance is calculated using monthly mean data with the mean seasonal cycle subtracted. Information on the calculation of error bars in **b** and **c** can be found in Methods.

the core and at the lower boundary of the EUC (Fig. 3b), and amplitudes of about 6 cm s^{-1} are derived at the surface. The variance explained by the 1,670-d harmonic oscillation decreases towards the surface (Fig. 3d), mainly as a result of the increasing strength of intraseasonal fluctuations. Although the vertical phase propagation is consistently downward below the EUC, the phase jumps by about 180° at the lower boundary of the EUC (Fig. 3c), approximately at the critical level for the propagation of high-baroclinic-mode equatorial Kelvin waves.

The 1,670-d fluctuations are also pronounced in subsurface temperature records. Temperatures are affected in two ways by the presence of equatorial deep jets: isopycnal displacements associated with the deep jets will lead to temperature variations that are phase-shifted in space and time relative to the velocity anomalies, depending on the character (Rossby or Kelvin) of the wave[10]; and in the presence of climatological zonal temperature gradients, zonal advection associated with the jets might induce changes in the temperature fields. For example, in-phase oscillations of surface zonal velocity and near-surface temperatures (Fig. 3c) are in agreement with the propagation of equatorial Kelvin waves; that is, eastward velocities are associated with downward isopycnal (isothermal) displacements and vice versa. A deeper thermocline could, in turn, be associated with reduced downward heat transport through diapycnal mixing causing higher SSTs. In the equatorial Atlantic, the climatological[27] zonal temperature gradient changes sign with depth, further complicating the interpretation of the observed phase structure of the subsurface temperature variability: for example, the reversal of the zonal temperature gradient with depth in the lower part of the EUC (Fig. 3a) might be responsible for the phase shift with depth of the 1,670-d harmonic oscillation of the subsurface temperature (Fig. 3c). Although the understanding of the propagation characteristics of the jets in the presence of strong mean currents and zonal tracer gradients deserves further theoretical study, these observations suggest that equatorial deep-jet energy propagates to the surface and affects sea surface conditions.

Observations in the equatorial Atlantic reveal a similar periodic behaviour for deep-jet oscillations over different time intervals and depth ranges[10,11]. Such consistent behaviour could arise from the development of high baroclinic basin modes[22] established by the eastward and westward propagation of Kelvin and Rossby waves, respectively[28]. In this case, vertical phase and energy propagation can occur only for quasi-resonant modes with active forcing and dissipation. The basin width of the Indian Ocean suggests a similar period for equatorial deep-jet oscillations as in the Atlantic, with rather different behaviour in the Pacific as a result of the much greater basin width. Argo float drift data from about 1,000-m depth represent a consistent data set that is available for all three oceans[19]. In the Atlantic, maximum

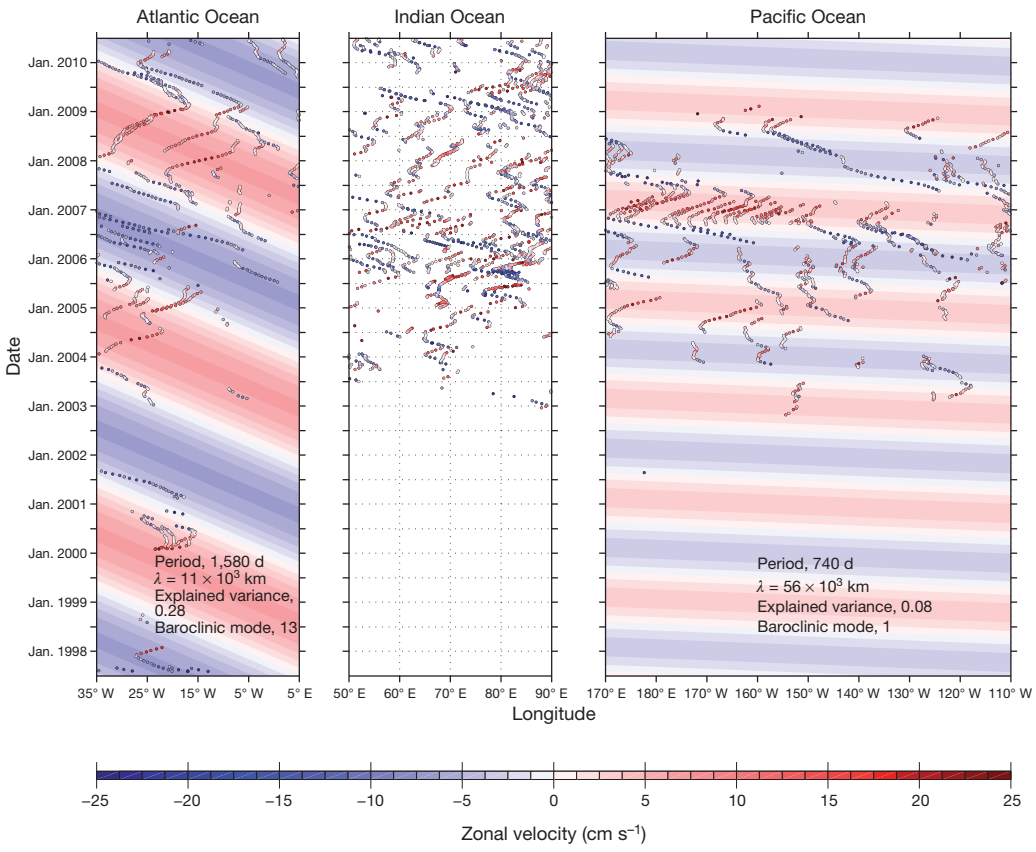


Figure 4 | Equatorial zonal velocities from 1,000-m Argo float drift data. Argo float drift data (coloured dots, colour scale) were acquired between latitudes 1° S and 1° N. The dominant interannual variability in the Atlantic and Pacific oceans obtained by maximizing explained variance using a plane-wave fit is visualized by colour shadings. Associated harmonic parameters for the Atlantic and Pacific oceans are given in the figure.

explained variance is found for westward-propagating Rossby waves of baroclinic mode 13 (corresponding to a vertical wavelength of about 600 m at 1,000-m depth) and a period of 1,580 d, corresponding to a zonal wavelength ($\lambda = 2\pi/|k|$, where k is the zonal wavenumber) of 11×10^3 km. In the Pacific, only weak signals of high-baroclinic-mode variability were extracted from the approximately 7-yr-long time series, which could be expected as estimated deep-jet oscillation periods are in the multidecadal range[12,20]. The dominant signal there is associated with low-baroclinic-mode variability. Despite there being geometric similarities between the Indian and Atlantic oceans, during the analysed time frame the Indian Ocean Argo float velocities are characterized by incoherent signals in the interannual period range, with no preferred period (Fig. 4). From this analysis, we expect no influence of equatorial deep jets on the surface conditions in the Indian and Pacific oceans on interannual timescales.

In analysing the seasonality of the Atlantic deep-jet surface expressions, we find that the amplitude of the 1,670-d cycle of zonal velocity is seasonally independent whereas the corresponding amplitudes of the ATL3 SST anomalies at this period are instead strongest during boreal summer and November/December (Supplementary Figs 6 and 7). These periods are identified as cold seasons with shallow thermocline depths in the east and active Bjerknes positive feedback[29,30]. During boreal spring when the tropical Atlantic is uniformly warm, the influence of the 1,670-d zonal velocity anomalies on SST is weak. Such behaviour is consistent with the equatorial zonal surface flow forced by interior ocean dynamics, whereas associated SST variations are seasonally modulated. On decadal timescales, the strength and period of the deep-jet oscillations may vary over time. The modulation could be due, for example, to a change in the dominant baroclinic mode affecting the basin mode period[22,28]. Such behaviour is suggested by Supplementary Fig. 8, although other modes of variability, such as the Pacific El Niño/Southern Oscillation[2] and the North Atlantic Oscillation[3], could also be influencing the time series. Despite this caveat, the surface expressions of the deep jets can clearly be used to improve the prediction of equatorial Atlantic SST, which is crucial for seasonal to interannual climate forecasting in the region[9].

METHODS SUMMARY

We calculated surface zonal velocity anomaly at the Equator, averaged between 35° W and 15° W (Fig. 1b), by applying a second-order fit in latitude to monthly mean meridional sea level anomaly distributions between 1° N and 1° S and evaluating equatorial geostrophy using the obtained curvature. The standard error of annual mean Argo float velocities (Fig. 1b) was calculated by dividing their standard deviation by the square root of the number of float observations. We filtered monthly time series (Fig. 1b) using a running annual mean. Harmonic analyses of zonal velocity and temperature (Figs 1b and 3) were performed by applying a linear regression model in a least-squares sense to the data. We approximated the degrees of freedom used for the calculation of the standard error of the resulting amplitudes as the length of the time series divided by a quarter of the deep-jet oscillation period. The significance of the correlation (Fig. 1a) was obtained using surface wind and rainfall time series of the same length as the microwave optimally interpolated SST with corresponding degrees of freedom. Sources and time intervals of all data sets used in this study are given in Supplementary Table 1.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 November 2010; accepted 21 March 2011.
Published online 18 May 2011.

1. Giannini, A., Saravanan, R. & Chang, P. Oceanic forcing of Sahel rainfall on interannual to interdecadal time scales. *Science* **302,** 1027–1030 (2003).
2. Chang, P., Fang, Y., Saravanan, R., Ji, L. & Seidel, H. The cause of the fragile relationship between the Pacific El Niño and the Atlantic Niño. *Nature* **443,** 324–328 (2006).
3. Czaja, A., van der Vaart, P. & Marshall, J. A diagnostic study of the role of remote forcing in tropical Atlantic variability. *J. Clim.* **15,** 3280–3290 (2002).
4. Carton, J. A., Cao, X., Giese, B. S. & da Silva, A. M. Decadal and interannual SST variability in the tropical Atlantic Ocean. *J. Phys. Oceanogr.* **26,** 1165–1175 (1996).
5. Chang, P., Ji, L. & Li, H. A decadal climate variation in the tropical Atlantic Ocean from thermodynamic air-sea interactions. *Nature* **385,** 516–518 (1997).
6. Zebiak, S. E. Air–sea interaction in the equatorial Atlantic region. *J. Clim.* **6,** 1567–1586 (1993).
7. Carton, J. A. & Huang, B. Warm events in the tropical Atlantic. *J. Phys. Oceanogr.* **24,** 888–903 (1994).
8. Chang, P. *et al.* Climate fluctuations of tropical coupled systems — The role of ocean dynamics. *J. Clim.* **19,** 5122–5174 (2006).
9. Kushnir, Y., Robinson, W. A., Chang, P. & Robertson, A. W. The physical basis for predicting Atlantic sector seasonal-to-interannual climate variability. *J. Clim.* **19,** 5949–5970 (2006).
10. Johnson, G. C. & Zhang, D. Structure of the Atlantic Ocean equatorial deep jets. *J. Phys. Oceanogr.* **33,** 600–609 (2003).
11. Bunge, L., Provost, C., Hua, B. L. & Kartavtseff, A. Variability at intermediate depths at the equator in the Atlantic Ocean in 2000–06: annual cycle, equatorial deep jets, and intraseasonal meridional velocity fluctuations. *J. Phys. Oceanogr.* **38,** 1794–1806 (2008).
12. Johnson, G. C., Kunze, E., McTaggart, K. E. & Moore, D. W. Temporal and spatial structure of the equatorial deep jets in the Pacific Ocean. *J. Phys. Oceanogr.* **32,** 3396–3407 (2002).
13. Luyten, J. R. & Swallow, J. C. Equatorial undercurrents. *Deep-Sea Res.* **23,** 999–1001 (1976).
14. Ruiz-Barradas, A., Carton, J. A. & Nigam, S. Structure of interannual-to-decadal climate variability in the tropical Atlantic sector. *J. Clim.* **13,** 3285–3297 (2000).
15. Wang, F. & Chang, P. A linear stability analysis of coupled tropical Atlantic variability. *J. Clim.* **21,** 2421–2436 (2008).
16. Ding, H., Keenlyside, N. S. & Latif, M. Equatorial Atlantic interannual variability: the role of heat content. *J. Geophys. Res.* **115,** C09020 (2010).
17. Jochum, M., Murtugudde, R., Malanotte-Rizzoli, P. & Busalacchi, A. in *Earth Climate: The Ocean-Atmosphere Interaction* (eds Wang, C., Xie, S.-P. & Carton, J. A.) 181–188 (Geophys. Monogr. Ser. 147, American Geophysical Union, 2004).
18. Philander, S. G. H. & Pacanowski, R. C. Response of equatorial oceans to periodic forcing. *J. Geophys. Res.* **86,** 1903–1916 (1981).
19. Lebedev, K. V., Yoshinari, H., Maximenko, N. A. & Hacker, P. W. *YoMaHa'07: Velocity Data Assessed from Trajectories of Argo Floats at Parking Level and at the Sea Surface.* IPRC Technical Note 4 (International Pacific Research Center, 2007).
20. Firing, E., Wijffels, S. E. & Hacker, P. Equatorial subthermocline currents across the Pacific. *J. Geophys. Res.* **103,** 21413–21423 (1998).
21. Ponte, R. M. & Luyten, J. R. Deep velocity measurements in the western equatorial Indian Ocean. *J. Phys. Oceanogr.* **20,** 44–52 (1990).
22. d'Orgeville, M., Hua, B. L. & Sasaki, H. Equatorial deep jets triggered by a large vertical scale variability within the western boundary layer. *J. Mar. Res.* **65,** 1–25 (2007).
23. Hua, B. L. *et al.* Destabilization of mixed Rossby gravity waves and the formation of equatorial zonal jets. *J. Fluid Mech.* **610,** 311–341 (2008).
24. Eden, C. & Dengler, M. Stacked jets in the deep equatorial Atlantic Ocean. *J. Geophys. Res.* **113,** C04003 (2008).
25. McPhaden, M. J., Proehl, J. A. & Rothstein, L. M. The interaction of equatorial Kelvin waves with realistically sheared zonal currents. *J. Phys. Oceanogr.* **16,** 1499–1515 (1986).
26. Brown, G. L. & Sutherland, B. R. Internal wave tunnelling through non-uniformly stratified shear flow. *Atmosphere-Ocean* **45,** 47–56 (2007).
27. Gouretski, V. V. & Koltermann, K. P. *WOCE Global Hydrographic Climatology.* Report 35 (Bundesamt für Seeschifffahrt und Hydrographie, 2004).
28. Cane, M. A. & Moore, D. W. A note on low-frequency equatorial basin modes. *J. Phys. Oceanogr.* **11,** 1578–1584 (1981).
29. Keenlyside, N. S. & Latif, M. Understanding equatorial Atlantic interannual variability. *J. Clim.* **20,** 131–142 (2007).
30. Okumura, Y. & Xie, S.-P. Some overlooked features of tropical Atlantic climate leading to a new Niño-like phenomenon. *J. Clim.* **19,** 5859–5874 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This study was supported by the German Federal Ministry of Education and Research as part of the co-operative project 'North Atlantic' and by the German Science Foundation as part of the Sonderforschungsbereich 754 'Climate-Biogeochemistry Interactions in the Tropical Ocean'. The contribution of J.M.T. was facilitated by support from the Woods Hole Oceanographic Institution's Columbus O'Donnell Iselin Chair for Excellence in Oceanography. We thank J. Fischer for mooring planning and field-work participation, F. Ascani for discussion and S.-H. Didwischus for data processing. This study uses PIRATA velocity and temperature data provided through the TAO project office, Argo float drift data provided by APDRC/IPRC[19], rainfall data from the Global Precipitation Climatology Project, Met Office Hadley Centre and microwave optimally interpolated SST data, NCEP/NCAR reanalysis wind data, and AVISO sea level anomaly data (Supplementary Table 1).

Author Contributions P.B. led the project and designed the study including sea-going work and data analysis. A.F. and V.H. processed and analysed moored velocity, Argo float and satellite data. J.M.T. performed moored profiler measurements, its data processing and its analysis. P.B., M.D. and R.J.G. led the drafting of the manuscript. All authors contributed to the interpretation of the results and provided substantial input to the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to P.B. (pbrandt@ifm-geomar.de).

METHODS

Surface zonal velocity anomaly at the Equator, averaged between 35° W and 15° W (Fig. 1b), was calculated by applying a second-order fit in latitude to monthly mean meridional sea level anomaly distributions between 1° S and 1° N and evaluating equatorial geostrophy using the obtained curvature. Mean zonal velocities from Argo float drifts between 950 and 1,050 m (Fig. 1b) were derived by removing outliers using a standard-deviation criterion and averaging over time (1-yr period) and space (from 1° S to 1° N and from 35° W to 15° W). The standard error of the nominal 1,000-m zonal velocities (Fig. 1b) was calculated by dividing their standard deviation by the square root of the number of float observations. Monthly time series (Fig. 1b) were filtered using a running annual mean. The dominant period of these time series was estimated by calculating the variance explained by a plane-wave fit (Supplementary Fig. 1).

In the subsurface temperature and velocity time series from PIRATA buoys and subsurface moorings, which are used to produce Fig. 3, data gaps are present. Here monthly time series were derived by monthly averaging and subtracting a mean annual cycle.

Harmonic analyses of zonal velocity and temperature time series (Figs 1b and 3b, c) were performed by fitting the following linear regression model in a least-squares sense to the monthly data:

$$\mathbf{d}_{\mathrm{m}} = g\boldsymbol{\beta} = \beta_1 \mathbf{I}_N + \beta_2 \cos(\omega \mathbf{t}) + \beta_3 \sin(\omega \mathbf{t})$$

Here $\mathbf{t}$ is the time vector corresponding to the data vector, $\mathbf{d}$, both of which are of length N; $\cos(\omega\mathbf{t})$ and $\sin(\omega\mathbf{t})$ are the vectors whose elements are the cosines and sines of the elements of $\omega\mathbf{t}$, respectively; $\omega = 2\pi/p$ is the angular frequency, where p is the period; g is the model matrix; $\boldsymbol{\beta}$ is a column vector of scalar model factors (β_1, β_2 and β_3); and $\mathbf{I}_N$ is a vector of length N whose elements all equal 1. The error matrix is given by

$$\Delta\beta = \sqrt{\frac{(g^{\mathrm{T}}g)^{-1}(\mathbf{d}-\mathbf{d}_{\mathrm{m}})^{\mathrm{T}}(\mathbf{d}-\mathbf{d}_{\mathrm{m}})}{n-k}}$$

where n is the number of degrees of freedom and $k = 2$ is the number of dependent model factors. The standard errors of the elements of $\boldsymbol{\beta}$ are given by the diagonal elements of $\Delta\beta$. The degrees of freedom used for the calculation of the standard error of the resulting amplitudes were approximated as the length of the time series divided by a quarter of the deep-jet oscillation period, and are $n = 14$ for ATL3 SST (HadISST), $n = 10$ for ATL3 SST (microwave optimally interpolated SST), $n = 14$ for geostrophic zonal velocity anomaly, $n = 10$ for the Argo float drift data (Fig. 1b and Supplementary Table 2), and $n = 5$ to $n = 7$ for the moored zonal velocity and subsurface temperature data (Fig. 3) varying with depth owing to data gaps. The phase errors (Fig. 3c) are maximum errors derived using the standard errors of the model factors ($\Delta\beta_2$ and $\Delta\beta_3$) by applying linear error propagation for an arbitrary phase lag.

The significance of the correlation (Fig. 1a) is obtained using surface wind and rainfall time series of the same length as the microwave optimally interpolated SST data series (Fig. 1b), which are additionally 270-d low-pass-filtered and have $n = 10$. Sources and periods of all data sets used are given in Supplementary Tab. 1.

Equatorial zonal velocities from 1,000-m Argo float drift data acquired between 1° S and 1° N were plotted as functions of time and longitude in Fig. 4. The data model

$$\mathbf{d}_{\mathrm{m}} = \hat{U}\sin(k\mathbf{x} - \omega\mathbf{t} - \phi\mathbf{I}_N)$$

was applied to the observed zonal velocities. Here $\hat{U}$ is the zonal velocity amplitude, $\sin(k\mathbf{x} - \omega\mathbf{t} - \phi\mathbf{I}_N)$ is the vector whose elements are the sines of the elements of $k\mathbf{x} - \omega\mathbf{t} - \phi\mathbf{I}_N$, $\mathbf{x}$ is the space vector in the zonal direction corresponding to the data vector, k is the zonal wavenumber and ϕ is the phase. By maximizing the variance explained by the fit, propagation characteristics of the dominant interannual variability were obtained (Supplementary Fig. 4). In the Atlantic, this fit explains about 28% of the variance of the equatorial zonal velocity from Argo float drift data after subtracting the annual and semi-annual cycles. In the Pacific, the strongest interannual signal (which explains only 8% of the variance) is found at a period of 740 d with a zonal wavelength of 56×10^3 km. The associated phase velocity corresponds to a first-baroclinic-mode Rossby wave that is very probably forced by the wind (Supplementary Fig. 4). Uncertainties in period and wavelength were estimated by a non-parametric bootstrap procedure where a number of resamples was constructed by random sampling with replacement (Supplementary Fig. 5).

Moored velocity data were acquired using acoustic Doppler current profilers, different single-point current meters and a moored profiler (Figs 2 and 3 and Supplementary Fig. 3). The oceanic variability on short timescales clearly exceeds the measurement accuracy of the different instruments. Owing to a ballasting error, the moored profiler was deployed 'light' and suffered loss of drive-wheel traction over time, resulting in truncation of the down-going profiles as time progressed (Fig. 2).

LETTER

doi:10.1038/nature09993

Inferring nonlinear mantle rheology from the shape of the Hawaiian swell

N. Asaadi[1], N. M. Ribe[2] & F. Sobouti[1]

The convective circulation generated within the Earth's mantle by buoyancy forces of thermal and compositional origin is intimately controlled by the rheology of the rocks that compose it. These can deform either by the diffusion of point defects (diffusion creep, with a linear relationship between strain rate and stress) or by the movement of intracrystalline dislocations (nonlinear dislocation creep)[1,2]. However, there is still no reliable map showing where in the mantle each of these mechanisms is dominant, and so it is important to identify regions where the operative mechanism can be inferred directly from surface geophysical observations. Here we identify a new observable quantity—the rate of downstream decay of the anomalous seafloor topography (swell) produced by a mantle plume—which depends only on the value of the exponent in the strain rate versus stress relationship that defines the difference between diffusion and dislocation creep. Comparison of the Hawaiian swell topography with the predictions of a simple fluid mechanical model shows that the swell shape is poorly explained by diffusion creep, and requires a dislocation creep rheology. The rheology predicted by the model is reasonably consistent with laboratory deformation data for both olivine[3] and clinopyroxene[4], suggesting that the source of Hawaiian lavas could contain either or both of these components.

Three distinct approaches have been used to constrain the rheological structure of the Earth's mantle. The most versatile approach is laboratory deformation experiments wherein the relation between the strain rate $\dot{\varepsilon}$ and the deviatoric stress τ in a deforming sample is measured under controlled physical and chemical conditions[1,2]. Such experiments yield rheological laws of the form $\dot{\varepsilon} = D\tau^n$, where n is a power-law exponent and D depends in general on pressure, temperature, grain size and the chemical composition and mineralogy of the sample. Diffusion creep has $n = 1$, whereas $n = 3.5$ for dislocation creep of olivine[5], the dominant mineral in the uppermost mantle. However, application of these laws to the Earth is subject to large uncertainties, both because several of the properties on which D depends are poorly constrained in the mantle and because geological strain rates are 6–8 orders of magnitude lower than those in the laboratory.

A second, geodynamical approach is to infer the depth variation of mantle viscosity from the rates of surface rebound following deglaciation events[6], the long-wavelength components of the Earth's nonhydrostatic gravitational equipotential surface (geoid)[7], or both together[8]. A robust result of this approach is that the lower mantle must be more viscous (by at least a factor of ten) than the upper mantle. However, studies using this approach typically assume a linear rheology ($n = 1$) throughout the mantle, and consequently provide little information about the relative importance of dislocation versus diffusion creep.

The third approach is seismological, and exploits the fact that dislocation creep (but not diffusion creep) causes the crystallographic axes of the crystals in a deforming rock to become aligned in a nonrandom way[9]. The resulting dependence of elastic wave speed on the propagation direction (seismic anisotropy) can be detected either by its effect on surface waves[10] or by the splitting of shear waves that it causes[11], and provides unambiguous evidence for active dislocation creep somewhere along the path between the source and receiver. However, the precise location of the anisotropic region along this path is difficult to constrain.

Here we show that a more direct determination of the operative deformation mechanism is possible in a well-defined portion of the Earth's uppermost mantle that lies beneath the Hawaiian islands. Since the earliest days of plate tectonic theory, the Hawaiian islands have inspired many fundamental new ideas in global geophysics, including the concept of a hotspot as a fixed locus of mantle melting[12,13], the association of hotspots with plumes rising from the lower mantle[13–15], the lithospheric flexure model of isostasy[16], and the recognition that hotspots can experience episodes of rapid migration[17]. Below we show that Hawaii can also make a novel contribution to our understanding of mantle rheology.

The largest geophysical signature produced by the Hawaiian plume is a broad topographic swell some 1,400 m high and 1,000 km wide (Fig. 1). According to the principle of gravitational equilibrium (isostasy), the downward buoyancy force on the swell (which displaces sea water) must be compensated by an upward buoyancy force provided by low-density material beneath: in this case, the anomalously hot material that ascends in the Hawaiian plume and then spreads laterally over the base of the Pacific plate.

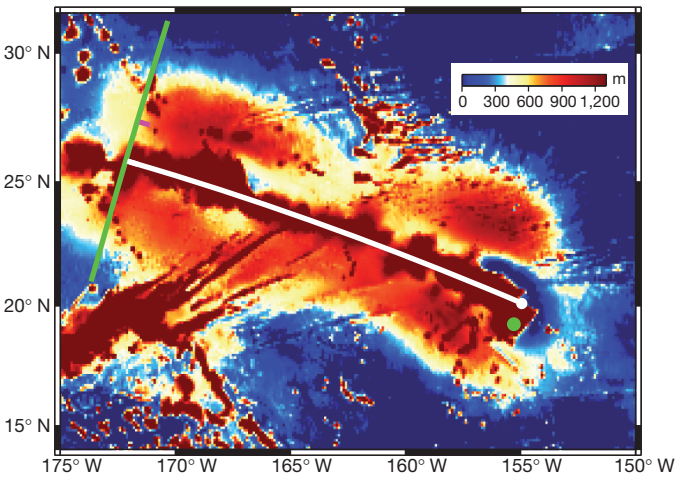


Figure 1 | Residual topography of the Hawaiian swell. This is defined as the observed seafloor topography[29] minus the reference topography predicted by a thermal model for cooling oceanic lithosphere[30]. Negative values of residual topography appear as dark blue, and values exceeding 1,300 m as red-brown. The white line is a segment of a great circle that fits the central axis of the island chain between longitudes 159–173° W. The green dot shows the current location of the centre of the Hawaiian hotspot (Kilauea volcano), and the white dot is its projection onto the central axis. Our analysis considers only the main part of the swell southeast of the green line.

[1]Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan 45137-66731, Iran. [2]Université Pierre et Marie Curie (Paris 6), Université Paris-Sud, CNRS, Lab FAST, Bâtiment 502, Campus Universitaire, 91405 Orsay, France.

Focusing now on the main part of the swell east of 173° W longitude, we note first that its width is smaller in the middle (around 163° W) than to either side, reflecting a past episode of reduced (by about 20%) upward flux of material in the Hawaiian plume[18]. However, the parts of the swell to the left of (older than) and the right of (younger than) the constriction have nearly the same amplitude and width. This crucial observation motivates the following physical model.

We consider the Hawaiian plume as a fixed source of buoyant fluid beneath a plate moving at constant speed[19,20] (Fig. 2a). The fluid spreads laterally over the plate's base to form a broad plume head with a thickness that is small compared to its lateral dimensions. Because the flow within it is dominantly simple shear on horizontal planes, the viscosity $\eta = \tau/\dot{\varepsilon}$ corresponding to the rheology $\dot{\varepsilon} = D\tau^n$ is:

$$\eta = D^{-\frac{1}{n}} \dot{\varepsilon}^{\frac{1-n}{n}} \tag{1}$$

where $\dot{\varepsilon} = \left|\frac{\partial \mathbf{u}}{\partial z}\right|$, $\mathbf{u}$ is the horizontal velocity and $\boldsymbol{z}$ is the vertical coordinate.

The thickness of a plume head with the rheology of equation (1) is governed by a partial differential equation that can be derived using the 'lubrication' theory of slow viscous flow in thin layers (Methods). It contains three parameters: the plate speed U, the volumetric supply rate Q of the buoyant material, and the parameter $\sigma = D(g\delta\rho)^n/[(n+2)(n+3)]$ that characterizes its rate of lateral spreading, where g is the gravitational acceleration and $-\delta\rho$ is the density deficit of the plume relative to the ambient mantle. A scaling analysis of the equation shows that the width of the plume head is proportional to:

$$L_0 = \left(\frac{\sigma Q^{2n+1}}{U^{2n+2}}\right)^{\frac{1}{3n+1}} \tag{2}$$

Conservation of the downstream volume flux then implies that the plume head's thickness is proportional to $h_0 = Q/(UL_0)$. Finally, isostasy implies that the amplitude of the topography is proportional to:

$$S_0 = \frac{Q\delta\rho}{UL_0(\rho_0 - \rho_w)} \tag{3}$$

where $\rho_0 = 3{,}400\,\mathrm{kg\,m^{-3}}$ is the density of the ambient mantle and $\rho_w = 1{,}000\,\mathrm{kg\,m^{-3}}$ that of sea water.

Figures 2b and c show the steady-state numerical solutions of the lubrication equation for diffusion creep ($n = 1$) and dislocation creep ($n = 3.5$) rheologies. Gravitational spreading is less efficient for a plume head with a dislocation creep rheology, which decays and widens more slowly downstream because the viscosity increases from the centre (where $\dot{\varepsilon}$ is largest) towards the edges (where $\dot{\varepsilon}$ is smaller). The influence of the rheology is revealed by a similarity solution of the lubrication equation that is valid at distances x far downstream from the hotspot (Methods). It predicts that the swell topography along the axis $y = 0$ should decay downstream as:

$$S \propto S_0 \left(\frac{x - x_0}{L_0}\right)^{-\frac{1}{3n+2}} \tag{4}$$

where x_0 is the virtual position of the hotspot. Accordingly, $S \propto (x - x_0)^{-0.2}$ for diffusion creep and $S \propto (x - x_0)^{-0.08}$ for dislocation creep with $n = 3.5$. The rate of downstream decay of the swell is therefore a sensitive indicator of the rheology of the buoyant material that compensates it.

To compare the predictions of the lubrication model with the observations, we first exclude the parts of the residual topography that are unrelated to the swell (Methods). Using a least-squares procedure, we determine the values of L_0 and S_0 for which the numerical solutions of Figs 2b and c best fit the non-excluded portion ($1.7 \times 10^6\,\mathrm{km^2}$) of the residual topography. These solutions are shown in Fig. 3a (for diffusion creep) and Fig. 3b (for dislocation creep), together with two contours (0 m and 700 m) of the residual topography. The diffusion creep solution does not match well the shape of the swell. It decays too rapidly downstream, and is too narrow near the hotspot and too wide farther downstream. In contrast, the dislocation creep solution matches much better the slow downstream decay of the swell and the shape of its boundary.

The above analysis can be refined by accounting for the variation of the Hawaiian plume flux Q during the past 20 million years. This gave rise to an along-chain variation of the swell's cross-sectional area $A(x)$, which has a minimum around 163° W longitude (Fig. 1). Figure 3 shows the solutions of the time-dependent lubrication equation for $n = 1$ and $n = 3.5$ that provide the best fit to the distribution $A(x)$ determined from the residual topography data[18] (Methods). Again, only the solution with $n = 3.5$ provides an acceptable fit to the residual topography.

We now consider the depth of the low-density material compensating the swell, which is not immediately obvious. An estimate is provided by the geoid–topography ratio (GTR), the ratio of the anomalous height of the gravitational equipotential surface to the topography anomaly. In the limit of a very broad swell, the GTR (in metres of geoid per kilometre of topography) is approximately one-tenth the average depth d (in kilometres) of the low-density compensating material. Early estimates of GTR ≈ 4–5 m km^{-1} for the Hawaiian swell[21] suggested $d \approx 40$–50 km, above the base of normal oceanic lithosphere of the same age (about 90 million years). However, those estimates were too low, owing to incomplete removal of the effect of the volcanic islands, which are more shallowly compensated (by the plate flexure mechanism) than the swell itself. Refined estimates that correct for this bias[22] yield GTR ≈ 7–8 m km^{-1}, indicating that the

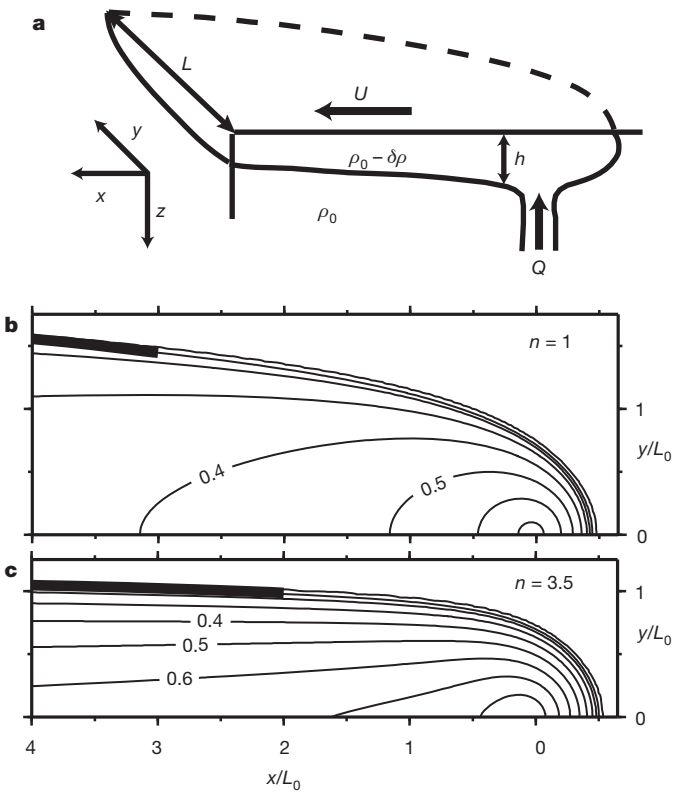


Figure 2 | Lubrication-theory model for the Hawaiian swell. **a**, Model geometry (oblique cut-away view). Buoyant fluid with a density deficit $-\delta\rho$ relative to the ambient mantle and a diffusion creep ($n = 1$) or dislocation creep ($n = 3.5$) rheology is released at a volumetric rate Q beneath a plate moving at speed U. The fluid spreads laterally to form a thin plume head of thickness h and half-width $L(x) \gg h$. **b**, Steady-state plume head thickness $h(x, y)$ predicted by the model for $n = 1$ (portion $y \geq 0$ only). The contour interval is $0.1Q/(UL_0)$, where L_0 is defined by equation (2). The heavy black line shows the width predicted by the analytical similarity solution (Methods). **c**, Same as **b**, but for $n = 3.5$.

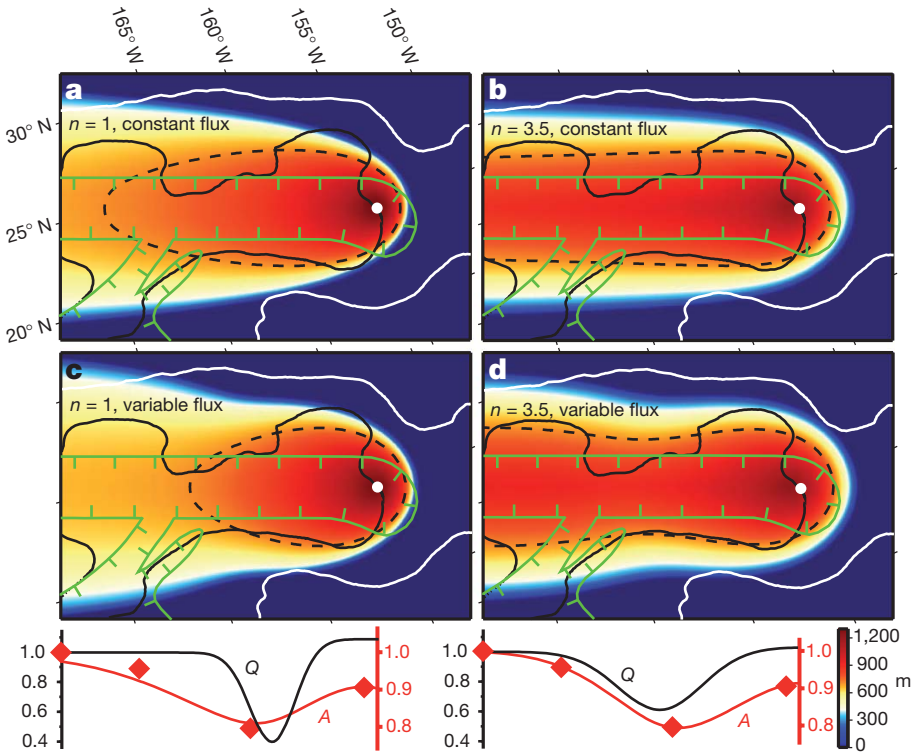


Figure 3 | Comparison of the lubrication model predictions with the residual topography of the Hawaiian swell. Data are shown as oblique Mercator projections. The solid lines are smoothed contours of the residual topography for 0 m (white) and 700 m (black). The green curve bounds the areas excluded from the analysis (Methods). Colours show the solutions of the lubrication equation for diffusion creep ($n = 1$; **a** and **c**), dislocation creep ($n = 3.5$; **b** and **d**), constant plume flux (**a** and **b**) and variable plume flux (**c** and **d**) that best fit the residual topography (Methods). The 700 m contour of each solution is shown by a dotted black line (compare with the solid black line). The flux $Q(t = -x/U)$ and the swell's cross-sectional area $A(x)$ for the two variable-flux cases are also shown below, each normalized to its value at the left of the figure. The diamond symbols are estimates (similarly normalized) of $A(x)$ obtained from the residual topography[18].

low-density material is at depths near or slightly above the base of normal lithosphere. This result, together with the well-defined lateral boundaries of the swell, implies that the region in which our inference of dislocation creep deformation applies is well constrained in all three dimensions. Our result is also consistent with geodynamical modelling of the apparent thermal age of Pacific lithosphere inferred from seismology, which suggests that dislocation creep is the dominant deformation mechanism at depths <410 km throughout the Pacific upper mantle[23].

We now use the numerical solution of Fig. 3d to estimate the values of the Hawaiian plume buoyancy flux $B \equiv Q\delta\rho$ and the rheological prefactor D in equation (1) (Methods). The best-fitting values of L_0 and S_0 for that solution imply $B = 5{,}610\,\mathrm{kg\,s^{-1}}$, within the range (2,800–8,700 $\mathrm{kg\,s^{-1}}$) of previous estimates[20,24,25]. The values also imply $D = (8.7\text{–}19.8) \times 10^{-38}\,\mathrm{kg\,m^{-1}\,s^{-12/7}}$, which is more easily understood by translating it into the minimum viscosity $\eta_{\min}$ within the plume head. For the solution of Fig. 3d and a realistic range of values of $\delta\rho$ (Methods), the maximum strain rate within the plume head is $\dot{\varepsilon}_{\max} = (3.1\text{–}3.7) \times 10^{-13}\,\mathrm{s^{-1}}$. Equation (1) then implies $\eta_{\min} = (2.3\text{–}3.3) \times 10^{19}$ Pa s.

The above viscosity estimate can be compared with laboratory-based rheological laws for dry[26,27] olivine aggregates as a function of temperature, pressure and strain rate[3] (Methods). For strain rates $(3.1\text{–}3.7) \times 10^{-13}\,\mathrm{s^{-1}}$ and representative values of temperature and pressure in the Hawaiian plume head, the experimental rheological law predicts a viscosity $\eta_{\exp} = (3.2\text{–}15) \times 10^{18}$ Pa s, a factor 1.6–10.3 lower than $\eta_{\min}$.

An alternative mineralogical model for the Hawaiian plume posits that a substantial portion of the source region of Hawaiian lavas is entirely free of olivine, being dominated instead by clinopyroxene[28]. It is therefore of interest to compare the viscosity inference from our model with the predictions of laboratory deformation experiments on natural clinopyroxene aggregates[4], for which $n = 4.7$. The solutions of the lubrication equation with $n = 4.7$ that best fit the Hawaiian residual topography are shown in Supplementary Fig. 3. The maximum strain rate is $\dot{\varepsilon}_{\max} = (6.3 - 7.6) \times 10^{-13}\,\mathrm{s^{-1}}$, which corresponds to a viscosity $\eta_{\min} = (1.0\text{–}1.5) \times 10^{19}$ Pa s. For the same values of temperature and pressure as previously, the experimental rheological law for clinopyroxene (Methods) predicts a viscosity $\eta_{\exp} = (3.4\text{–}7.5) \times 10^{18}$ Pa s, a factor of 1.3–4.4 lower than $\eta_{\min}$. Our model predictions thus agree slightly better with the laboratory data for clinopyroxene than with those for olivine. However, the error bars are large because the applicability of laboratory-based rheological laws to mantle conditions is difficult to demonstrate. We conclude that our inferred rheology is probably consistent with a source comprising either olivine-rich or clinopyroxene-rich components, or both.

METHODS SUMMARY

The lubrication model with constant σ was validated against a three-dimensional numerical code for convection in a fluid with viscosity dependent on temperature, pressure and strain rate (Supplementary Figs 1 and 2). The lubrication equation was solved numerically using an explicit finite-difference method. The scales L_0 and S_0 that yield the best fit of a given dimensionless solution of the lubrication equation to the residual topography data were determined by maximizing the variance reduction. For non-steady-state solutions, the time-varying flux $Q(t)$ was also adjusted iteratively to obtain the best possible fit between the cross-sectional area distribution $A(x)$ of the swell predicted by the lubrication solution and that estimated from the observations.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 14 September 2010; accepted 7 March 2011.
Published online 11 May 2011.

1. Poirier, J.-P. *Creep of Crystals* (Cambridge University Press, 1985).
2. Karato, S.-I. *Deformation of Earth Materials* (Cambridge University Press, 2008).

3. Keefner, J. W., Mackwell, S. J., Kohlstedt, D. L. & Heidelbach, F. Dependence of the creep of dunite on oxygen fugacity: implications for viscosity variations in Earth's mantle. *J. Geophys. Res.* doi:10.1029/2010JB00748 (in the press).
4. Bystricky, M. & Mackwell, S. Creep of dry clinopyroxene aggregates. *J. Geophys. Res.* **106,** 13443–13454 (2001).
5. Bai, Q., Mackwell, S. & Kohlstedt, D. L. High-temperature creep of olivine single crystals. 1. Mechanical results for buffered samples. *J. Geophys. Res.* **96,** 2441–2463 (1991).
6. Peltier, W. R. Glacial isostatic adjustment. 2: Inverse problem. *Geophys. J. R. Astron. Soc.* **46,** 669–705 (1976).
7. Hager, B. H., Clayton, R. W., Richards, M. A., Comer, R. P. & Dziewonski, A. M. Lower mantle heterogeneity, dynamic topography and the geoid. *Nature* **313,** 541–545 (1985).
8. Mitrovica, J. X. & Forte, A. M. A new inference of mantle viscosity based upon joint inversion of convection and glacial isostatic adjustment data. *Earth Planet. Sci. Lett.* **225,** 177–189 (2004).
9. Karato, S. Jung. H., Katayama, I. & Skemer, P. Geodynamic significance of seismic anisotropy of the upper mantle: new insights from laboratory studies. *Annu. Rev. Earth Planet. Sci.* **36,** 59–95 (2008).
10. Montagner, J. P. & Tanimoto, T. Global upper mantle tomography of seismic velocities and anisotropies. *J. Geophys. Res.* **96,** 20337–20351 (1991).
11. Long, M. D. & Silver, P. G. Shear wave splitting and mantle anisotropy: measurements, interpretations, and new directions. *Surv. Geophys.* **30,** 407–461 (2009).
12. Wilson, J. T. A possible origin of the Hawaiian islands. *Can. J. Phys.* **41,** 863–870 (1963).
13. Morgan, W. J. Convection plumes in the lower mantle. *Nature* **230,** 42–43 (1971).
14. Montelli, R. *et al.* Finite-frequency tomography reveals a variety of plumes in the mantle. *Science* **303,** 338–343 (2004).
15. Wolfe, C. J. *et al.* Mantle shear-wave velocity structure beneath the Hawaiian hot spot. *Science* **326,** 1388–1390 (2009).
16. Watts, A. B. & Cochran, J. R. Gravity anomalies and flexure of the lithosphere along the Hawaiian-Emperor Seamount Chain. *Geophys. J. R. Astron. Soc.* **38,** 119–141 (1974).
17. Tarduno, J. A. *et al.* The Emperor seamounts: southward motion of the Hawaiian hotspot plume in Earth's mantle. *Science* **301,** 1064–1069 (2003).
18. Davies, G. F. Temporal variation of the Hawaiian plume flux. *Earth Planet. Sci. Lett.* **113,** 277–286 (1992).
19. Olson, P. in *Magma Transport and Storage* (ed. Ryan, M.) 33–51 (John Wiley, 1990).
20. Ribe, N. M. & Christensen, U. Three-dimensional modelling of plume-lithosphere interaction. *J. Geophys. Res.* **99,** 669–682 (1994).
21. Marks, K. M. & Sandwell, D. T. Analysis of geoid height versus topography for oceanic plateaus and swells using nonbiased linear regression. *J. Geophys. Res.* **96,** 8045–8055 (1991).
22. Cserepes, L., Christensen, U. & Ribe, N. M. Geoid height versus topography for a plume model of the Hawaiian swell. *Earth Planet. Sci. Lett.* **178,** 29–38 (2000).
23. van Hunen, J., Zhong, S., Shapiro, N. M. & Ritzwoller, M. H. New evidence for dislocation creep from 3-D geodynamic modeling of the Pacific upper mantle structure. *Earth Planet. Sci. Lett.* **238,** 146–155 (2005).
24. Sleep, N. H. Hotspots and mantle plumes: some phenomenology. *J. Geophys. Res.* **95,** 6715–6736 (1990).
25. Ribe, N. M. & Christensen, U. The dynamical origin of Hawaiian volcanism. *Earth Planet. Sci. Lett.* **171,** 517–531 (1999).
26. Hirth, G. & Kohlstedt, D. L. Water in the oceanic upper mantle: implications for rheology, melt extraction and the evolution of the lithosphere. *Earth Planet. Sci. Lett.* **144,** 93–108 (1996).
27. Karato, S.-I. Insights into the nature of plume-asthenosphere interaction from central Pacific geophysical anomalies. *Earth Planet. Sci. Lett.* **274,** 234–240 (2008).
28. Sobolev, A. V., Hofmann, A. W., Sobolev, S. V. & Nikogosian, I. K. An olivine-free mantle source of Hawaiian shield basalts. *Nature* **434,** 590–597 (2005).
29. Smith, W. H. F. & Sandwell, D. T. Global seafloor topography from satellite altimetry and ship depth soundings. *Science* **277,** 1956–1962 (1997).
30. Stein, C. & Stein, S. A model for the global variation in oceanic depth and heat flow with lithospheric age. *Nature* **359,** 123–129 (1992).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Davaille, C. Herzberg, S.-I. Karato and D. Kohlstedt for discussions and advice. This work was supported by the French embassy in Tehran and by the SEDIT programme of INSU and the ANR (grant PTECTO) in France.

Author Contributions N.A. derived the lubrication equation, determined the similarity solution and the full numerical solutions of that equation, and analysed the topography data. N.M.R. proposed the idea for the study, determined the three-dimensional numerical solutions with temperature-dependent rheology, and wrote the manuscript. F.S. co-directed the parts of the work done in Zanjan. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to N.A. (n_asaadi@iasbs.ac.ir) or N.M.R. (ribe@fast.u-psud.fr).

METHODS

Derivation and solution of the lubrication equation. In the following, the subscripts α and β take on the values 1 or 2, corresponding to vector components in the x- and y-directions, respectively.

Because the plume head is thin, the pressure inside it is nearly hydrostatic. Its value relative to the pressure outside is:

$$p = g\delta\rho(h-z) \tag{5}$$

where g is the gravitational acceleration and $-\delta\rho$ is the density deficit of the plume relative to the ambient mantle. The horizontal flow in the plume head is a Poiseuille flow driven by the lateral gradient of the pressure. In the lubrication approximation, the horizontal velocity components u_α satisfy:

$$\partial_z(\eta\partial_z u_\alpha) = \partial_\alpha p \equiv g\delta\rho\partial_\alpha h \tag{6}$$

where the viscosity η is given by equation (1). The solution of equation (6) for which the slip $u_\alpha - U\delta_{1\alpha}$ at the plume head's upper surface $z = 0$ and the shear traction $\eta\partial_z u_\alpha$ at its lower surface $z = h$ both vanish is:

$$u_\alpha = U\delta_{1\alpha} - \frac{D}{n+1}(g\delta\rho)^n|\nabla h|^{n-1}\left[h^{n+1} - (h-z)^{n+1}\right]\partial_\alpha h \tag{7}$$

where $\delta_{1\alpha}$ is the Kronecker delta symbol.

Conservation of mass over the whole thickness of the plume head requires:

$$\partial_t h + \partial_\beta \int_0^h u_\beta \mathrm{d}z = \frac{Q}{\pi a^2}\exp\left(-r^2/a^2\right) \tag{8}$$

where the source of buoyant fluid has been represented by a radially symmetric Gaussian distribution of width a with $r^2 = x^2 + y^2$, and summation over the repeated subscript β is understood. Substitution of equation (7) into equation (8) finally yields:

$$\partial_t h + U\partial_x h = \nabla\cdot\left[\sigma|\nabla h|^{n-1}\nabla\left(h^{n+3}\right)\right] + \frac{Q}{\pi a^2}\exp\left(-r^2/a^2\right) \tag{9}$$

where $\sigma = D(g\delta\rho)^n/[(n+2)(n+3)]$.

Equation (9) was prepared for numerical solution by rewriting it in terms of the dimensionless variables hUL_0/Q, x/L_0, y/L_0 and tU/L_0. Solutions were obtained using an explicit finite-difference method. The gravitational spreading term (first term on the right-hand side) was discretized using a centred difference scheme with fourth-order accuracy. The advection term $U\partial_x h$ was discretized using second-order accurate upwind differencing. Time stepping was performed explicitly using a second-order accurate midpoint method.

Far downstream from the hotspot, the source term in equation (9) is negligible, and gravitational spreading is primarily in the y direction. The steady-state form of equation (9) then takes the form:

$$U\partial_x h = \sigma(n+3)\partial_y\left(h^{n+2}\left|\partial_y h\right|^{n-1}\partial_y h\right) \tag{10}$$

where σ has been assumed constant. Conservation of the downstream volume flux requires:

$$U\int_{-L}^{L} h\,\mathrm{d}y = Q \tag{11}$$

where $L(x)$ is the half-width of the plume head. Equations (10) and (11) admit a similarity solution of the form:

$$h = \frac{Q}{UL(x)}H(\zeta) \tag{12}$$

where $\zeta = \dfrac{y}{L(x)}$. The appropriate boundary conditions are:

$$0 = L(x_0) = H'(0) = H(1) \tag{13}$$

where x_0 is the virtual position of the hotspot as seen by the solution far downstream and the prime indicates $\mathrm{d}/\mathrm{d}\zeta$. Substitution of equation (12) into equations (10) and (11) yields ordinary differential equations for $L(x)$ and $H(\zeta)$ that can be solved subject to the conditions (13) to yield:

$$H(\zeta) = c_1\left(1 - |\zeta|^{\frac{1+n}{n}}\right)^{\frac{n}{2n+1}} \tag{14}$$

and

$$L = c_2 L_0\left(\frac{x - x_0}{L_0}\right)^{\frac{1}{3n+2}} \tag{15}$$

where

$$c_1 = \frac{\Gamma\left(\frac{5n^2+5n+1}{(n+1)(2n+1)}\right)}{2\Gamma\left(\frac{2n+1}{n+1}\right)\Gamma\left(\frac{3n+1}{2n+1}\right)} \tag{16}$$

$$c_2 = \left[(n+3)(3n+2)\left(\frac{n+1}{2n+1}\right)^n c_1^{2n+1}\right]^{\frac{1}{3n+2}} \tag{17}$$

and Γ is the Gamma function. For $n = 3.5$, $c_1 \approx 0.6775$ and $c_2 \approx 0.9432$. The virtual hotspot position x_0 is determined by least-squares fitting of equation (15) to the full numerical solution of equation (9), yielding $x_0/L_0 = 0.753$ for $n = 1$ and 0.272 for $n = 3.5$.

Validation of the lubrication model. Although the gravitational spreading parameter σ depends on temperature and pressure in general, we treat it as constant in this study. Here we justify this assumption using a three-dimensional numerical code[20,31] for convection in a fluid whose viscosity η depends on temperature, pressure and strain rate according to:

$$\eta = \eta_{\min} + \left(\eta_{\max}^{-1} + \eta_1^{-1}\right)^{-1}, \tag{18}$$

$$\eta_1 = \eta_0\left(\frac{\dot{\varepsilon}_1}{\dot{\varepsilon}_0}\right)^{\frac{1-n}{n}}\exp\left(\frac{E+pV}{nRT}\right), \tag{19}$$

$$\dot{\varepsilon}_1 = \left(I^2 + I_{\min}^2\right)^{1/2} \tag{20}$$

where $I = \left(2e_{ij}e_{ij}\right)^{1/2}$ is the second invariant of the strain rate tensor e_{ij}. In equations (18) to (20), $\eta_0 = 10^{21}$ Pa s is a reference viscosity, and $\eta_{\min} = 0.001\eta_0$ and $\eta_{\max} = 500\eta_0$ are the minimum and maximum allowable viscosities, respectively. $\dot{\varepsilon}_0 = 2\times10^{-15}\ \mathrm{s}^{-1}$ is a reference strain rate, and $I_{\min} = 10^{-15}\ \mathrm{s}^{-1}$ is the minimum allowable value of I. Finally, T is the absolute temperature, p is the pressure, E is the activation energy, V is the activation volume, and R is the universal gas constant.

The equations of conservation of mass, momentum and energy are solved using a hybrid spectral (in the two horizontal directions) and finite-difference (in the vertical direction) method in which the coupling of different spectral modes by lateral viscosity variations is calculated iteratively[20,31]. A thermal plume is generated by imposing a hot temperature anomaly $\Delta T_{\max}\exp(-r^2/a^2)$ on the bottom of the model box (depth $d = 400$ km). The thermal plume interacts with the shear flow generated by motion of the upper surface of the model box at a constant velocity $U = 2.7\times10^{-9}\ \mathrm{m\,s^{-1}}$ (8.6 cm yr^{-1}). Effects of melting-induced depletion[25] are neglected for simplicity.

Supplementary Fig. 1a shows the steady-state temperature field for a plume with buoyancy flux $B = 2{,}340\ \mathrm{kg\,s^{-1}}$ and $\Delta T_{\max} = 200$ K in a fluid with diffusion creep ($n = 1$) rheology with $E = 400\ \mathrm{kJ\,mol^{-1}}$ and $V = 6.1\times10^{-6}\ \mathrm{m^3\,mol^{-1}}$. To compare this three-dimensional solution quantitatively with the lubrication model, we use the former to calculate the isostatic topography:

$$S_{\mathrm{iso}} = \frac{\rho_0\alpha}{\rho_0 - \rho_w}\int_0^d \delta T\ \mathrm{d}z \tag{21}$$

where $\delta T(x, y, z)$ is the temperature anomaly (temperature with the plume present less the temperature in a second solution with $\Delta T_{\max} = 0$) and $\alpha = 3.5\times10^{-5}\ \mathrm{K^{-1}}$ is the coefficient of thermal expansion. Supplementary Fig. 1b shows S_{iso} on the central axis $y = 0$ as a function of $x - x_0$ with $x_0 = 163$ km (black line), together with the power-law decay of equation (4) with $n = 1$ that best fits it (red line). The downstream decay of the topography agrees closely with the prediction $S_{\mathrm{iso}} \propto (x - x_0)^{-0.2}$ of the lubrication model.

Supplementary Fig. 2 is for a plume with the same values of B, $\Delta T_{\max}$, E and V as in Supplementary Fig. 1, but with a dislocation creep rheology ($n = 3.5$). Again, the topography decays downstream in accordance with the lubrication prediction $S_{\mathrm{iso}} \propto (x - x_0)^{-0.08}$.

Supplementary Figs 1 and 2 show that the lubrication model accurately reproduces the dynamics of the more complicated three-dimensional code for both diffusion creep ($n = 1$) and dislocation creep ($n = 3.5$) rheologies, even in the presence of strongly temperature-dependent viscosity. The physical reason for this is that the largest temperature variations in the three-dimensional code are confined near the upwelling plume conduit, so that lateral variations of temperature throughout most of the plume head are relatively small (Supplementary Figs 1a and 2a). A lubrication model with constant σ can therefore be used instead of the full three-dimensional code, which, because of its high computational cost and multiplicity of parameters, would render intractable the task of finding the model with the optimal fit to the residual topography data.

Comparison of the model predictions with the residual topography. The first step is to exclude those portions of the residual topography that are unrelated to the swell, which include the Hawaiian islands, the surrounding moat produced by lithospheric flexure, and the nearby Mid-Pacific mountains. These areas are enclosed by the green lines in Fig. 3. We also excluded all topography outside the range 0–1,300 m.

Once a numerical solution of equation (9) was found for a given value of n, we used a least-squares procedure (maximization of variance reduction) to determine the values L_0^{bf} and S_0^{bf} of the length scale L_0 and the topography scale S_0 such that the re-dimensionalized solution best fits the non-excluded residual topography. For a steady-state solution with constant Q, the determination is direct. For non-steady-state solutions in which Q is a function of time, we determined L_0^{bf} and S_0^{bf} iteratively by adjusting the shape of the function $Q(t)$ to obtain the best possible agreement between the cross-sectional area distribution $A(x)$ of the swell topography predicted by the numerical model and that estimated directly from the residual topography data[18]. An iterative procedure is required because the volume flux $Q(t)$ of the Hawaiian plume varies on a timescale that is comparable to the fluid-mechanical adjustment time L_0/U, which implies that the cross-sectional area of the swell at a given distance x from the hotspot is not simply proportional to the plume flux at the earlier time $t = -x/U$ when the cross-section was directly over the hotspot. For the non-steady lubrication solution of Fig. 3d, $L_0^{bf} = 630$ km and $S_0^{bf} = 1{,}375$ m.

Given L_0^{bf} and S_0^{bf}, the buoyancy flux B and the rheological prefactor D of the best-fitting lubrication solution were found by solving the two simultaneous equations $L_0 = L_0^{bf}$ and $S_0 = S_0^{bf}$, where L_0 and S_0 are defined by equations (2) and (3) respectively. The results are:

$$B = S_0^{bf} L_0^{bf} (\rho_0 - \rho_w) U \tag{22}$$

$$D = \frac{(2+n)(3+n)\left(L_0^{bf}/g\right)^n \delta\rho^{1+n} U}{\left[S_0^{bf}(\rho_0 - \rho_w)\right]^{1+2n}} \tag{23}$$

To evaluate equations (22) and (23) we assumed $U = 2.7 \times 10^{-9}\,\mathrm{m\,s^{-1}}$ and $\delta\rho = \rho_0 \alpha \Delta T$, where $\alpha = 3.5 \times 10^{-5}\,\mathrm{K^{-1}}$ and $\Delta T = 125$–150 K is a typical excess temperature in the plume head far from the hotspot, based on a maximum temperature contrast (250–300 K) of the upwelling material beneath the hotspot[32].

Comparison of the model predictions with laboratory experiments. Laboratory deformation experiments in the dislocation creep regime typically yield rheological laws of the form:

$$\dot{\varepsilon} = \mathcal{A} f^m \tau^n \exp\left(-\tfrac{E+pV}{RT}\right) \tag{24}$$

where $\mathcal{A}$ is a constant and f is the oxygen fugacity. To evaluate equation (24) within the Hawaiian plume head downstream from the hotspot, we assume $p = 3.3$ GPa (corresponding to a depth $\approx$100 km) and $T = 1{,}748$–$1{,}773$ K, corresponding to an excess temperature $\Delta T = 125$–150 K relative to an ambient mantle temperature 1,350°C. For dry olivine aggregates, $m = 0.2$, $n = 3.59$, $\mathcal{A} = 1.15 \times 10^{-19}\,\mathrm{s^{-1}\,Pa^{-n-m}}$, $E = 449\,\mathrm{kJ\,mol^{-1}}$ (ref. 3), and $V = 1.7$–$2.8 \times 10^{-5}\,\mathrm{m^3\,mol^{-1}}$ (refs 33 and 34). In addition, we assume $f \in [10^{-6.2}, 10^{-2.1}]$ Pa (ref. 3). For dry natural clinopyroxene aggregates, $m = 0$, $n = 4.7$, $\mathcal{A} = 4.0 \times 10^{19}\,\mathrm{s^{-1}\,Pa^{-n}}$, and $E + pV = 760\,\mathrm{kJ\,mol^{-1}}$ for an average confining pressure $P = 425$ MPa (ref. 4). Because V is poorly known for clinopyroxene, we assume the same range of values as for olivine.

31. Christensen, U. & Harder, H. 3-D convection with variable viscosity. *Geophys. J. Int.* **104,** 213–226 (1991).
32. Herzberg, C. & Asimow, P. D. Petrology of some oceanic island basalts: PRIMELT2.XLS software for primary magma calculation. *Geochem. Geophys. Geosyst.* **9,** Q09001 (2008).
33. Kirby, S. H. Rheology of the lithosphere. *Rev. Geophys.* **21,** 1458–1487 (1983).
34. Borch, R. S. & Green, H. W. II. Deformation of peridotite at high pressure in a new molten salt cell: comparison of traditional and homologous temperature treatments. *Phys. Earth Planet. Inter.* **55,** 269–276 (1989).

LETTER

doi:10.1038/nature09943

Earth's earliest non-marine eukaryotes

Paul K. Strother[1], Leila Battison[2], Martin D. Brasier[2] & Charles H. Wellman[3]

The existence of a terrestrial Precambrian (more than 542 Myr ago) biota has been largely inferred from indirect chemical and geological evidence associated with palaeosols[1,2], the weathering of clay minerals[3] and microbially induced sedimentary structures in siliciclastic sediments[4]. Direct evidence of fossils within rocks of non-marine origin in the Precambrian is exceedingly rare[5,6]. The most widely cited example comprises a single report of morphologically simple mineralized tubes and spheres interpreted as cyanobacteria, obtained from 1,200-Myr-old palaeokarst in Arizona[5]. Organic-walled microfossils were first described from the non-marine Torridonian (1.2–1.0 Gyr ago) sequence of northwest Scotland in 1907[7]. Subsequent studies[8–10] found few distinctive taxa—a century later, the Torridonian microflora is still being characterized as primarily nondescript "leiospheres"[11]. We have comprehensively sampled grey shales and phosphatic nodules throughout the Torridonian sequence. Here we report the recovery of large populations of diverse organic-walled microfossils extracted by acid maceration, complemented by studies using thin sections of phosphatic nodules that yield exceptionally detailed three-dimensional preservation. These assemblages contain multicellular structures, complex-walled cysts, asymmetric organic structures, and dorsiventral, compressed organic thalli, some approaching one millimetre in diameter. They offer direct evidence of eukaryotes living in freshwater aquatic and subaerially exposed habitats during the Proterozoic era. The apparent dominance of eukaryotes in non-marine settings by 1 Gyr ago indicates that eukaryotic evolution on land may have commenced far earlier than previously thought.

The Torridonian is a thick (up to 12 km) sequence of immature siliciclastic rocks deposited in three unconformable Groups: Stoer, Sleat and Torridon (Supplementary Figs 1, 2). The Stoer Group (Pb–Pb age, 1,199 ± 70 Myr (ref. 12); ^{40}Ar–^{39}Ar age, 1,178.6 ± 9 Myr (ref. 13)) is unconformably overlain by the Torridon Group, which has been dated as old as 994 ± 48 Myr on the basis of Rb–Sr isochrons in the lowermost Diabaig Formation[12]. The Sleat Group has not been dated, but it is conformably overlain with strata correlated with the Torridon Group: the Applecross Formation is present, and parts of the Kinlock Formation have been correlated with the Diabaig Formation.

The Torridonian rocks consist largely of compositionally immature, coarse-grained, siliciclastic redbeds with lesser red and grey shales. Numerous lines of lithological evidence point towards a non-marine depositional setting for the entire Torridonian sequence. These include evidence of valley-confined alluvial fans, rivers and unconfined bajadas in the Stoer Group[14]; braided rivers[15] and fan-delta/lake deposits in the Sleat Group[14–16]; and, in addition to these fluvial and alluvial deposits, valley-confined lakes in the Torridon Group[14]. The interpretation of the grey shales of the Diabaig Formation (the basal unit of the Torridon Group) as lacustrine deposits is reinforced by a low boron content[17], small wave ripples (Supplementary Fig. 3) and pervasive desiccation cracks[4,14] (Supplementary Fig. 4), in combination with a lack of marine features such as tidal bundles and evaporites accompanying desiccation. Raindrop impressions found at Upper Diabaig and at Point Stoer provide additional evidence of subaerial exposure of mudrock units (Supplementary Fig. 5). It has been concluded[4] that sedimentary structures in the Diabaig Formation preserve non-marine microbial mats. Recent regional studies of the Torridonian, incorporating zircon provenance, continue to support a non-marine depositional setting for the entire sequence[11].

Samples of grey shale were collected from 17 different localities (Supplementary Figs 1, 2, Supplementary Table 1), of which 11 sections were reasonably productive (Supplementary Table 2). All of our samples, with the exception of TOR08-9, an isolated sample from Tarskavaig on Sleat, are from previously published stratigraphic sections where an inferred depositional environment based on geological grounds had been determined[15]. Dark grey shales, from which all palynomorphs were extracted, are considered to be lake bottom sediments, although at Cailleach Head, some of the yellowish grey mudstones are interpreted to be delta toe sediments[15]. These sediments retain the allochthonous character of normal sedimentation—they formed when rivers and streams scoured the regional surface and transported those sediments into a basin. The palynological content of our samples represents a combination of this allochthonous input plus *in situ* microfossils. Palynological samples were supplemented by thin section studies of phosphatic nodules from Loch Diabaig and Cailleach Head, which support the largely allochthonous nature of these deposits because they do not retain evidence of *in situ* microbial biofabrics.

Palynological assemblages (see Methods Summary) are dominated by simple spherical palynomorphs (sphaeromorph acritarchs) without distinctive surface ornament or sculpture, of the kind usually placed in the genus *Leiosphaeridia* (Fig. 1a). Surface ornament, when present, is limited to low verrucae (Fig. 1b) or scattered granae (Fig. 1c, d). Acanthomorph (with spines) acritarchs have not been observed. Excystment features include simple median splits (Fig. 1e) and circular pylomes (Fig. 1f). Such features, including pre-formed sutures in the cell wall, are associated with a eukaryotic level of cellular structure[18]. The appearance and texture of vesicle walls is remarkably varied, reflecting primary differences in thickness, construction and original composition. Some vesicle walls are structurally complex, as in Fig. 1g, h, and Supplementary Fig. 6. Here the vesicle wall is composed of roughly parallel, short cylindrical subunits, which impart a coarsely corrugate appearance in surface (Fig. 1g) view and a beaded appearance equatorially (Fig. 1h, Supplementary Fig. 6a, e). Overall, this multicellular fossil (Fig. 1h, Supplementary Fig. 6a, e) consists of a mass of cells tightly packed together and encased within a spherical vesicle. The bluntly ellipsoidal specimen in Fig. 1i has a distinctive wall, characterized by a highly-ordered arrangement of circular pits creating a reticulate pattern (Fig. 1j). This undescribed taxon superficially resembles *Dictyosphaera*, but lacks the plate-like wall structure of this Mesoproterozoic eukaryote[19]. Cell clusters are quite common, including *Synsphaeridium* spp. (Fig. 1k), *Chlorogloeaopsis* spp. (German) Hofmann and *Torridoniphycus lepidus* Zhang (in part). Many cell clusters show features such as mutually adpressed walls (Fig. 2a), indicating that these are not simply random associations of solitary cells. The preservation of internal bodies within vesicle walls is quite common; such a morphological feature is seen in numerous other Proterozoic

[1]Department of Earth and Environmental Sciences, Boston College, Weston, Massachusetts 02493, USA. [2]Department of Earth Sciences, University of Oxford, Parks Road, Oxford OX1 3PR, UK. [3]Department of Animal & Plant Sciences, The University of Sheffield, Sheffield S10 2TN, UK.

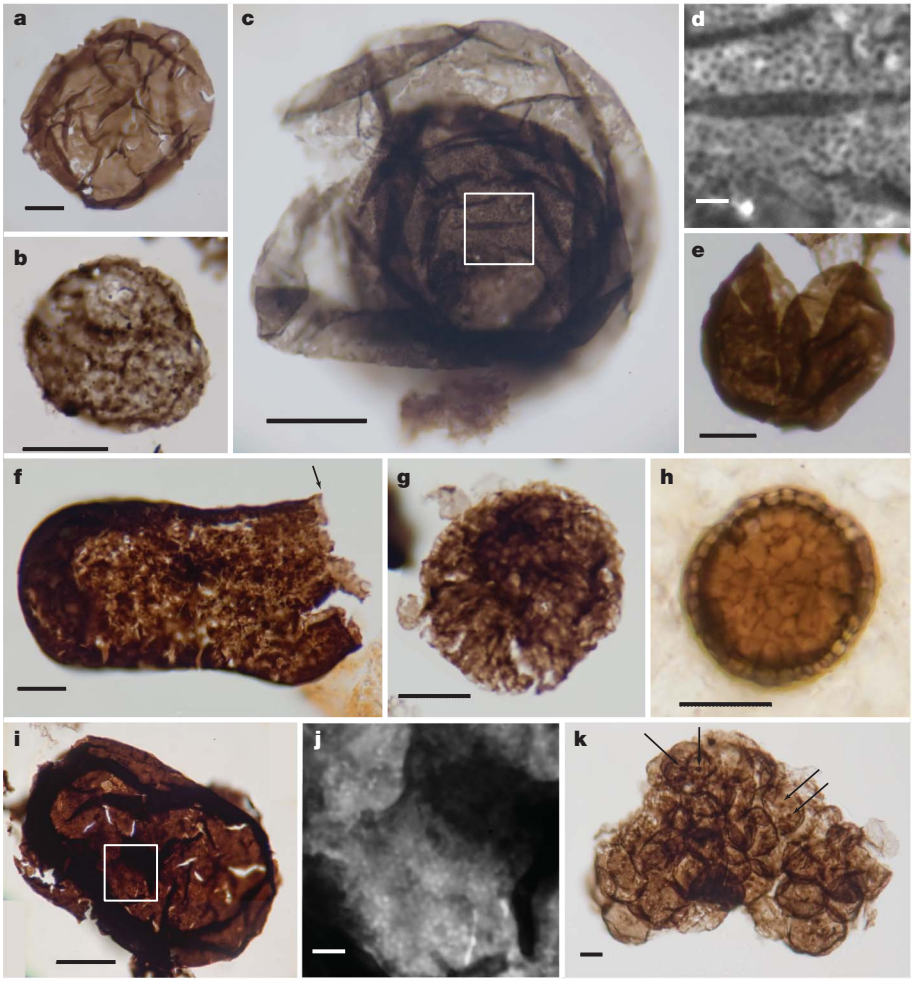


Figure 1 | Sphaeromorph acritarchs and cell clusters from the Torridonian, NW Scotland. **a**, *Leiosphaeridia crassa*; TOR08-18/Glame Member, Applecross Formation. **b**, Acritarch similar to *Trematoligotriletum emarginatum* Tim. with irregular verrucate surface; TOR08-26/Allt na Beistre Member, Applecross Formation. **c**, *Lophosphaeridium* sp. enclosed in a thin membranous vesicle; TOR08-34/Diabaig Formation. **d**, Surface detail of inner cyst (*Lophosphaeridium* sp.) in **c** showing small, evenly distributed granae. **e**, *Leiosphaeridia crassa* exhibiting a median split; TOR08-45/Cailleach Head Formation. **f**, Ellipsoidal cyst with granular wall structure exhibiting a terminal circular pylome excystment feature (arrow); TOR08-34/Diabaig Formation. **g**, Coarsely corrugate vesicle with dense contents; TOR08-27/Diabaig Formation. **h**, Spherical ball of cells enclosed within a complex wall (thin section from phosphatic nodule, Diabaig Formation). **i**, Blunt ellipsoidal vesicle with a micro-reticulate wall; TOR08-46/Cailleach Head Formation. **j**, Detail of **i** (box), showing the reticulate wall texture. **k**, Cell cluster, similar to *Synsphaeridium* sp. Note included condensed organic 'spots' (arrows); TOR08-26/Allt na Beistre Member, Applecross Formation. Scale bars: 10 μm (**a**–**c**, **e**–**i**, **k**), 1 μm (**d**, **j**).

assemblages[6,20]. These dense organic 'spots' occur in both macerated specimens (Fig. 1k, arrows) and thin sections that retain three-dimensional preservation (Fig. 2a, arrows).

The Torridonian assemblages contain some striking examples of microfossils that show complexity that goes considerably beyond that of simple leiospheres. Figure 1h illustrates a multicellular sphere from a phosphatic nodule with a clearly differentiated outer wall that is similar to the dispersed corrugated forms in Fig. 1g and Supplementary Fig. 6e, f. A confocal laser scanning image of the same specimen (Supplementary Fig. 6a), representing a 0.2-μm-thick slice through the equatorial plane of the vesicle sphere, reveals a solid mass of mutually compressed cells. Some of these interior cells (Fig. 1h) retain a dense 'spot', probably the plasmolysed remnants of original cell contents. Figure 2b illustrates a large fusiform vesicle (475 μm wide) with a pitted (Fig. 2c) interior wall structure. Figure 2d shows a dark, heterogeneous central body (cb) enclosed within a large cyst with a peripheral asymmetrical structure (as), which itself appears to consist of several cylindrical cells or membranous outgrowths of the vesicle wall (Fig. 2e). The enclosed central body does not appear multicellular when viewed in transmitted infrared light (Fig. 2f). Another example of a large vesicle with asymmetric features is illustrated in Fig. 2g. This specimen lacks complex internal features, but does contain a single degraded central body (cb) and a clearly differentiated asymmetric structure (as) composed of several stubby projections.

Not all the fossils recovered are vesicular in their gross organization. The specimen in Fig. 3a possesses two arm-like projections (arm) with bluntly rounded tips attached to a large elliptical disk. One of the arms appears to be constricted at its basal attachment site (ba), giving the impression of either multicellular or coenocytic organization. This fossil is unlike any previously described acritarch. The tri-lobate thalloid form illustrated in Fig. 3b is 915 μm wide and represents the largest intact fossil recovered to date from the Torridonian sequence. The thallus is preserved as a dense, thick, organic layer, the upper surface of which appears as a dense cuticle-like protective layer. Internally the thallus appears irregularly reticulate when viewed in transmitted infrared light (Fig. 3c). This could be a reflection of an underlying parenchymatous cell structure which is now considerably degraded, and we interpret this specimen to be an example of biological structure that was approaching a tissue-level grade of organization. There is no evidence that the thallus was composed of filaments or has retained an underlying filamentous structure which might be construed as evidence for fungal or lichen affinity. Several examples of poorly preserved, spine-like structures

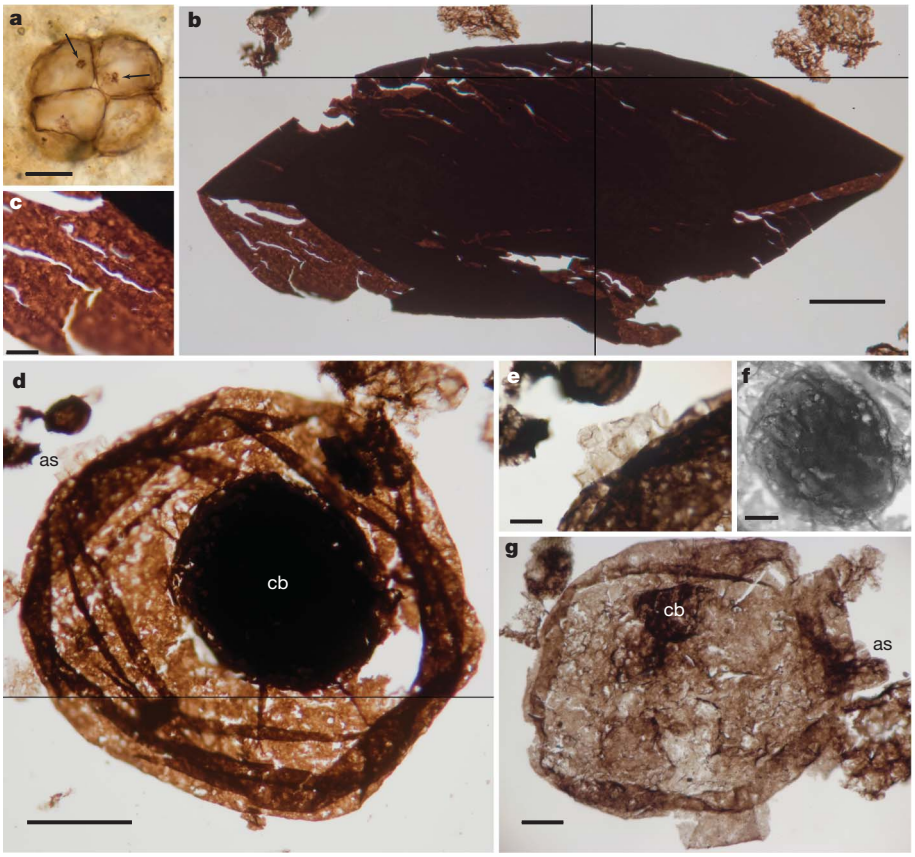


Figure 2 | Cell clusters and large, morphologically complex vesicles. Lines in **b** and **d** demarcate edges of separate images used to construct the photo mosaics. **a**, Cell cluster exhibiting mutually adpressed cells with included 'spots' (arrowed). This image is a photomontage of three different focal planes (thin section from phosphatic nodule, Diabaig Formation). **b**, Large fusiform vesicle with pitted/reticulate wall structure; TOR08-9b/Kinloch Formation. **c**, Detail from **b**, showing a portion of the inside of the vesicle wall. **d**, Large vesicle with a dense central body (cb) and an asymmetric structure (as); TOR08-34/Diabaig Formation. **e**, Detail of the asymmetric structure in **d**. **f**, Transmitted infrared (>830 nm) image of the central body in **c**, showing it to be a thick-walled, probably unicellular cyst. **g**, Large vesicle with a single-layered inner cyst and a large asymmetric structure (as) which appear as stubby projections from the main vesicle wall; TOR08-12/Kinloch Formation. Scale bars: 10 μm (**a**, **c**, **e**, **f**); 50 μm (**b**); 25 μm (**d**, **g**).

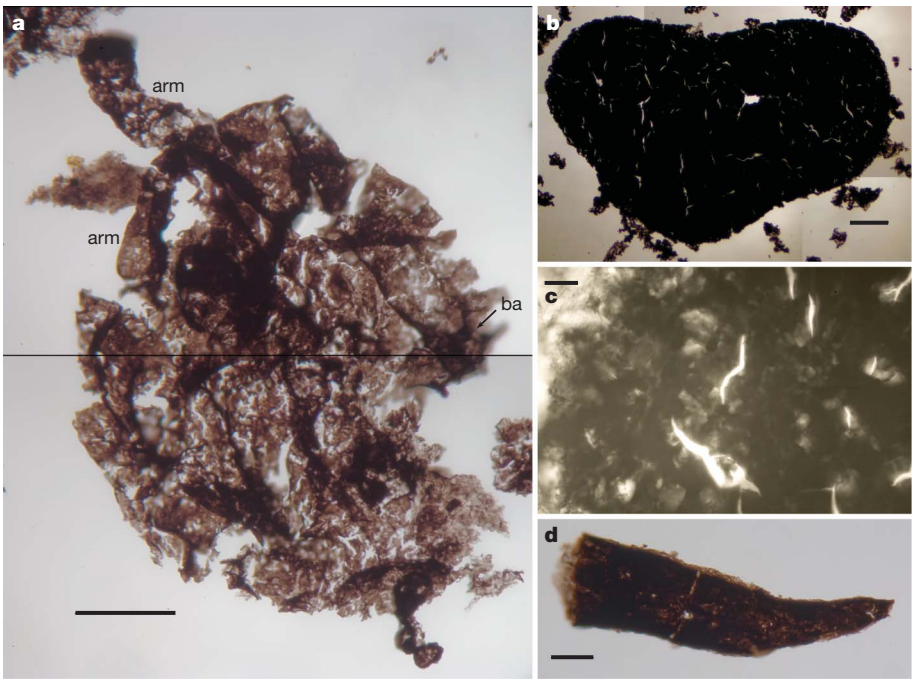


Figure 3 | Non-vesicular organic structures. **a**, Oval plate with two blunt-tipped arms (arm), one of which is attached (ba, arrowed) with what appears to be a basal plug; TOR08-32/Diabaig Formation. **b**, Tri-lobed thalloid organism with a dense upper surface showing small cracks and a heterogeneous inner layer. Image is a photomontage of four photographs; TOR08-40/Diabaig Formation. **c**, Detail of **b** photographed in infrared (>830 nm) transmitted light to reveal the internal structure. **d**, Spine (appendage?) tip; TOR08-34/Diabaig Formation. Scale bars: 50 μm (**a**); 100 μm (**b**); 10 μm (**c**, **d**).

were recovered (Fig. 3d). None resemble the spines or processes typically associated with acanthomorphic acritarchs.

A set of morphological criteria has been put forward[18,21] for assessing the record of early eukaryotes recovered in marine habitats by 1,500 Myr ago. Similar arguments can applied to the interpretation of microfossils found in non-marine settings at 1 Gyr ago. The excellent preservation in the Torridonian deposits has permitted the retention of putative vegetative stages; this enables a greater range of form and biological structure to be recognized. In Fig. 1c, a thicker-walled cyst which is ornamented with uniformly distributed granae (Fig. 1d, *Lophosphaeridium* sp.) is enclosed within a thinner-walled vesicle interpreted to be the original vegetative cell wall. This combination of two different wall types and the sculptured inner wall appears to be a eukaryotic feature. The possession of a pre-determined excystment opening, such as a median split (Fig. 1e), has been taken to indicate cytoskeletal control of excystment consistent with eukaryotic complexity[18]. This is reinforced in Fig. 1f (arrow), which shows a preformed circular opening which would have functioned in excystment.

Several larger microfossils also possess asymmetric features, which appear to be autapomorphies associated with evolution of unicellular eukaryotes. One large specimen (Fig. 2d) from the Sleat Group (Kinloch Formation) sports a singular set of thin-walled cylindrical structures (as) attached at one place on the surface (Fig. 2e). The enclosed dense central body (cb) appears vesicular and cyst-like (Fig. 2f), implying that the outer vesicle was originally vegetative. The folded arms in Fig. 3a arguably required cytoskeletal control of growth of the kind associated with a eukaryotic level of cellular organization. This combination of large size, topology, variable wall texture, and wall structure indicates that most of the Torridonian microfossils recovered from maceration are likely to have been of eukaryotic origin.

Taxonomic overlap between sphaeromorph acritarchs recognized here in lacustrine settings (Supplementary Table 2) and near-shore marine settings elsewhere in the Neoproterozoic[6,21–23] does not necessarily indicate that non-marine acritarch species are a common component of near-shore marine assemblages. Genera such as *Leiosphaeridia*, *Synsphaeridium* and *Lophosphaeridium* may be too morphologically simple to assure meaningful biological homology between samples of different ages and depositional settings.

The preservation of a three-dimensional ball of cells enclosed by a complex cell wall (Fig. 1h, Supplementary Fig. 6a–d) corresponds to a level of complexity that is not typical of prokaryotes. On the other hand, the enclosed cells are only about 2 μm in diameter, seemingly too small (and too ancient) to represent a metazoan blastula. Intriguingly, these seemingly multicellular structures may be compared with the early palintomic phase that forms part of the synzoospore hypothesis for metazoan origins[24], during which the generative cell (oöcyte) of a unicellular protoctist cleaves internally to produce smaller cells which remain attached to each other (synzoospores), forming a multicellular ball of cells. The simplicity of these balls of cells precludes their systematic assignment within the Eukarya. However, their morphology, in combination with larger, probably multicellular thalli (Fig. 3 b), indicates that evolutionary processes that preceded tissue-grade multicellularity in marine settings[25], such as cell-to-cell adhesion, were also evident in non-marine settings by 1 Gyr ago.

Sample to sample heterogeneity seen throughout the Torridonian (Supplementary Table 2, Supplementary Fig. 7) clearly indicates a significant degree of biotic diversity, reflecting adaptation to freshwater aquatic and subaerially exposed habitats by earliest Neoproterozoic time. Early eukaryotes were clearly capable of diversifying within non-marine habitats, not just in marine settings as has been generally assumed. This idea directly supports phylogenomic studies which find that the cyanobacteria evolved first in freshwater habitats and later migrated into marine settings[26]. Our findings also lend support to recent inferences regarding the impact of Neoproterozoic terrestrial biotas on Earth's biogeochemical cycles[3,13,27–29]. Freshwater habitats are ecologically more variable than marine habitats[30], providing temporal and physiochemical heterogeneity, including wet-drying cycles and direct atmosphere–organism gas exchange. Such habitat heterogeneity translates directly into increased speciation potential. Some of the microfossils illustrated here must have experienced subaerial exposure because they occur *in situ* in microbially induced sedimentary structures with desiccation cracks (Supplementary Fig. 4), but the extent to which they lived subaerially cannot be ascertained with certainty. Even so, gross morphology, in combination with an apparent lack of planktonic adaptive morphology, in the form of processes or spines, strongly suggests that a range of benthic, freshwater habitats were already colonized by eukaryotes by the beginning of the Neoproterozoic era.

METHODS SUMMARY

Palynological samples were prepared using conventional acid maceration techniques. Following HCl-HF-HCl acid maceration, the residues were sieved using a 10 μm mesh. They were then treated to a heavy liquid separation using zinc chloride, followed by further sieving at 10 μm. The organic residues were mounting directly onto glass slides using epoxy resin.

Phosphatic nodules were studied using petrographic thin sections cut parallel to bedding and crossing bedding. These were ground to varying thicknesses to optimize transparency of the dark phosphate and the position of the fossils within the phosphate.

Microscopical analysis was undertaken using transmitted white light supplemented by infrared analysis and laser confocal microscopy. All photomicrographs were recorded as 16 bit raw files (3,043 × 2,036 pixels) using a FujiFilm S5 IS Pro (infrared) digital camera body attached to a Zeiss Universal microscope equipped with Zeiss PlanApo 63× and Zeiss Plan-Neofluor 25× objectives. White light photomicrographs were photographed through a B+W 486 ultraviolet/infrared blocking filter to achieve a visible light colour balance; infrared images were captured using a long pass filter (830 nm, Edmund Scientific). Images were captured with FujiFilm Studio Utility software running on an iMac. An approximately neutral grey background was achieved by setting white to 219 in the levels menu in Photoshop. The image in Supplementary Fig. 6a was obtained with a Leica SP5 confocal laser scanning microscope using ultraviolet excitation at 405 nm and collection between 418 and 780 nm. Except where noted, no sharpening or any other image processing was used. Infrared images were converted to greyscale from RGB using the channel mixer adjustment menu in Photoshop CS4.

Received 13 October 2010; accepted 16 February 2011.
Published online 13 April 2011.

1. Ohmoto, H. Evidence in pre-2.2 Ga paleosols for the early evolution of atmospheric oxygen and terrestrial biota. *Geology* **24,** 1135–1138 (1996).
2. Gutzmer, J. & Beukes, N. J. Earliest laterites and possible evidence for terrestrial vegetation in the Early Proterozoic. *Geology* **26,** 263–266 (1998).
3. Kennedy, M., Droser, M., Mayer, L. M., Pevear, D. & Mrofka, D. Late Precambrian oxygenation; inception of the clay mineral factory. *Science* **311,** 1446–1449 (2006).
4. Prave, A. R. Life on land in the Proterozoic: evidence from the Torridonian rocks of northwest Scotland. *Geology* **30,** 811–814 (2002).
5. Horodyski, R. J. & Knauth, L. P. Life on land in the Precambrian. *Science* **263,** 494–498 (1994).
6. Schopf, J. W. & Klein, C. *The Proterozoic Biosphere* (Cambridge Univ. Press, 1992).
7. Teall, J. J. H. in *The Geological Structure of the North-west Highlands of Scotland* (eds Peach, B. N. *et al.*) 288, plate LII (Memoirs of the Geological Society of Great Britain, 1907).
8. Downie, C. So-called spores from the Torridonian. *Proc. Geol. Soc. Lond.* **1600,** 127–128 (1962).
9. Cloud, P. E. & Germs, A. New pre-Paleozoic nannofossils from the Stoer Formation (Torridonian), NW Scotland. *Geol. Soc. Am. Bull.* **82,** 3469–3474 (1971).
10. Zhang, Z. Upper Proterozoic microfossils from the Summer Isles, N.W. Scotland. *Palaeontology* **25,** 443–460 (1982).
11. Kinnaird, T. C. *et al.* The late Mesoproterozoic-early Neoproterozoic tectonostratigraphic evolution of NW Scotland: the Torridonian revisited. *J. Geol. Soc. Lond.* **164,** 541–551 (2007).
12. Turnbull, M. J. M., Whitehouse, M. J. & Moorbath, S. New isotopic age determinations for the Torridonian, NW Scotland. *J. Geol. Soc. Lond.* **153,** 955–964 (1996).
13. Parnell, J., Boyce, A. J., Mark, D., Bowden, S. & Spinks, S. Early oxygenation of the terrestrial environment during the Mesoproterozoic. *Nature* **468,** 290–293 (2010).
14. Stewart, A. D. *The Later Proterozoic Torridonian Rocks of Scotland: Their Sedimentology, Geochemistry and Origin* (Memoirs of the Geological Society, no. 24, 2002).
15. Stewart, A. D. Greywacke sedimentation in the Torridonian of Colonsay and Oronsay. *Geol. Mag.* **99,** 399–419 (1962).

16. Sutton, J. & Watson, J. Sedimentary structures in the epidotic grits of Skye. *Geol. Mag.* **97,** 106–122 (1960).
17. Stewart, A. D. & Parker, A. Palaeosalinity and environmental interpretation of red beds from the late Precambrian ('Torridonian') of Scotland. *Sedim. Geol.* **22,** 229–241 (1979).
18. Knoll, A. H., Javaux, E. J., Hewitt, D. & Cohen, P. Eukaryotic organisms in Proterozoic oceans. *Phil. Trans. R. Soc. B* **361,** 1023–1038 (2006).
19. Meng, F., Zhou, C., Yin, L., Chen, Z. & Yuan, X. The oldest known dinoflagellates: morphological and molecular evidence from Mesoproterozoic rocks at Yongji, Shanxi Province. *Chin. Sci. Bull.* **50,** 1230–1234 (2005).
20. Jankauskas, T. V., Mikhailova, N. S. & Hermann, T. N. in *Mikrofossilii Dokembriya SSSR [Precambrian Microfossils of the USSR]* 190 (Nauka, Leningrad, 1989).
21. Javaux, E. J., Knoll, A. H. & Walter, M. R. Morphological and ecological complexity in early eukaryotic ecosystems. *Nature* **412,** 66–69 (2001).
22. Knoll, A. H. Microbiotas of the late Precambrian Hunnberg Formation, Nordaustlandet, Svalbard. *J. Paleontol.* **58,** 131–162 (1984).
23. Butterfield, N. J. & Chandler, F. W. Palaeoenvironmental distribution of Proterozoic microfossils, with an example from the Agu Bay Formation, Baffin Island. *Palaeontology* **35,** 943–957 (1992).
24. Mikhailov, K. V. *et al.* The origin of Metazoa: a transition from temporal to spatial cell differentiation. *Bioessays* **31,** 758–768 (2009).
25. Butterfield, N. J. Modes of pre-Ediacaran multicellularity. *Precambr. Res.* **173,** 201–211 (2009).
26. Blank, C. E. & Sánchez-Baracalo, P. Timing of morphological and ecological innovations in the cyanobacteria – a key to understanding the rise in atmospheric oxygen. *Geobiology* **8,** 1–23 (2010).
27. Lenton, T. M. & Watson, A. J. Biotic enhancement of weathering, atmospheric oxygen and carbon dioxide in the Neoproterozoic. *Geophys. Res. Lett.* **31,** 1–5 (2004).
28. Knauth, L. P. & Kennedy, M. J. The late Precambrian greening of the Earth. *Nature* **460,** 728–732 (2009).
29. Spinks, S. C., Parnell, J. & Bowden, S. A. Reduction spots in the Mesoproterozoic age: implications for life in the early terrestrial record. *Int. J. Astrobiol.* **9,** 209–216 (2010).
30. Hutchinson, G. E. The paradox of the plankton. *Am. Nat.* **95,** 137–145 (1961).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements J. Rosenberg produced the confocal laser scanning image (Supplementary Fig. 6a); we thank O. Green for preparation of phosphatic nodules at Oxford. We thank J. Antcliffe, R. Callow and S. Moorhouse for field assistance and the people of Scoraig and Bill (the boatman) for access to Cailleach Head. This research was supported by NASA grant NNX07AU79G (P.K.S.), NERC NE/G015716/1 (C.H.W.) and NERC NE/G524060/1 (L.B.).

Author Contributions All authors contributed to the intellectual content, design and writing of the manuscript, and collection and study of phosphatic nodules. C.H.W. and P.K.S. collected the palynological samples. P.K.S. wrote an initial draft, prepared the photographic plates and produced the provisional taxonomic assessment. L.B. and C.H.W. drafted the figures.

Author Information All materials (rock sample, remaining organic residues, palynological slides, thin sections) are curated in the collections of the Centre for Palynology of the University of Sheffield, UK. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to P.K.S. (Strother@bc.edu).

LETTER

doi:10.1038/nature09920

A giant Ordovician anomalocaridid

Peter Van Roy[1,2] & Derek E. G. Briggs[1,3]

Anomalocaridids, giant lightly sclerotized invertebrate predators, occur in a number of exceptionally preserved early and middle Cambrian (542–501 million years ago) biotas and have come to symbolize the unfamiliar morphologies displayed by stem organisms in faunas of the Burgess Shale type. They are characterized by a pair of anterior, segmented appendages, a circlet of plates around the mouth, and an elongate segmented trunk lacking true tergites with a pair of flexible lateral lobes per segment[1,2]. Disarticulated body parts, such as the anterior appendages and oral circlet, had been assigned to a range of taxonomic groups—but the discovery of complete specimens from the middle Cambrian Burgess Shale showed that these disparate elements all belong to a single kind of animal[3]. Phylogenetic analyses support a position of anomalocaridids in the arthropod stem, as a sister group to the euarthropods[4–6]. The anomalocaridids were the largest animals in Cambrian communities. The youngest unequivocal examples occur in the middle Cambrian Marjum Formation of Utah[7] but an arthropod retaining some anomalocaridid characteristics is present in the Devonian of Germany[5]. Here we report the post-Cambrian occurrence of anomalocaridids, from the Early Ordovician (488–472 million years ago) Fezouata Biota[8] in southeastern Morocco, including specimens larger than any in Cambrian biotas. These giant animals were an important element of some marine communities for about 30 million years longer than previously realized. The Moroccan specimens confirm the presence of a dorsal array of flexible blades attached to a transverse rachis on the trunk segments; these blades probably functioned as gills.

Anomalocaridids were first described from the middle Cambrian of Mount Stephen in British Columbia, Canada[9]. The specimens, which consisted of isolated raptorial appendages, were interpreted as the body of a shrimp[9]. A different raptorial appendage from the middle Cambrian Burgess Shale was subsequently misinterpreted by Walcott as that of the arthropod *Sidneyia*[3,10]. Walcott interpreted the oral circlet as a medusoid and also described an incomplete specimen of the body as a holothurian[11]. The carapace elements of the anomalocaridid *Hurdia* also were misidentified originally—as two different arthropods[6,12,13]. The first reconstruction of an anomalocaridid based on complete specimens dates from the 1980s[2,3]. As a result of discoveries from other Cambrian localities worldwide, it was largely accepted by the mid-1990s that anomalocaridids are arthropods[1]. Some authors, however, continued to argue for non-arthropodan affinities[14] and the exact systematic placement of the group remained uncertain. Although inclusion of anomalocaridids in the arthropod crown has been favoured by some[15], they are generally regarded as belonging in the euarthropod stem[4–6,16].

Anomalocaridids are known from early and middle Cambrian sites in Canada, the United States, Poland, Russia, China, Australia and, possibly, Greenland and the Czech Republic[6,7,16–20]. The oldest known example is *Cassubia infercambriensis* from the early Cambrian of Poland[16], while the youngest are from the middle Cambrian of the United States[7]; a single specimen from the late Cambrian of the Holy Cross Mountains in Poland may represent an oral circlet[21]. It has been suggested that the Early Ordovician *Pseudoangustidontus duplospineus* might represent an anomalocaridid spine[22] or a complete grasping appendage[20]. Its morphology, however, differs from an anomalocaridid spine[20], and its unsegmented, sclerotized nature is inconsistent with an anterior appendage[22]. *Schinderhannes bartelsi* from the Lower Devonian Hunsrück Slate retains some anomalocaridid characters, including a circular mouth and anterior raptorial appendages, but clearly lies crownward as a sister taxon to the euarthropods[5].

The Fezouata Biota preserves a fully marine Burgess Shale fauna in combination with more advanced taxa[8,23]. The anomalocaridid specimens are associated with a diverse fauna[8] and were recovered from five excavations north of Zagora, in southeastern Morocco. Excavations 1 and 2 are approximately coeval, and sit just below the boundary between the Lower and Upper Fezouata Formations; they are latest Tremadocian in age. The remaining excavations are slightly younger, but all three occupy approximately the same stratigraphic level: they sit at the base of the Upper Fezouata Formation, dating them to the earliest Floian (Supplementary Fig. 1).

The Fezouata anomalocaridids occur in two distinct preservational styles. Smaller specimens, like the majority of Fezouata fossils[8,23], are flattened in the mudstones and preserved in pyrite weathered to iron oxides. Larger specimens occur in massive concretions, dominated by authigenic silica, in close proximity to each other within the mudstones. The composition of the fossils indicates weathered pyrite, but a significant proportion of manganese is present. Concretion formation was probably the result of the rapid decay of a large amount of organic material (large trilobites are also preserved in this way[24]). The concretions provided a degree of protection from compaction, so that the trunk retains some convexity and the lateral lobes project at a high angle.

The large anomalocaridid specimens were collected at Excavation 1. They comprise articulated bodies but they lack most or all of the head region. The most complete specimen (Fig. 1a–c, Supplementary Fig. 2) preserves the posterior margin of the head and a trunk of 11 segments; there is no evidence of tergites. The maximum preserved length and width (excluding lateral lobes) of this fossil (Fig. 1a) are respectively ~915 mm and ~295 mm. Specimens YPM 226438 and YPM 226439 (Supplementary Fig. 3a, b) are less complete, but their trunks attain a comparable width, indicating that they were of similar size. The segments are approximately equal in length except for the posteriormost 3 or 4, which become progressively shorter where the trunk tapers to a blunt end. They are covered by long, flexible blade-like structures, more than 100 in the widest segments, oriented parallel or slightly oblique to the axis of the trunk (Fig. 1a, b, Supplementary Figs 2, 3a). They are interpreted as covering the dorsal surface of the segments, based on the position of similar structures in *Laggania*[14,25,26]. Their attitude may be straight to gently curved or even sinuous; they run parallel to each other with limited overlap, except where they are folded and distorted in the incomplete terminal body segment of YPM 226437 (Supplementary Fig. 2c, d). There is no evidence that the blades continue laterally across the small triangular, flexible lobes that project from the ventro-lateral margin of the trunk (Fig. 1a–c; Supplementary Figs 2a, b, 3a, b); their exact relationship to the base of the lobes is unclear. Only the most proximal part of the lobes is clearly preserved; narrow closely spaced rays run parallel to the axis (Fig. 1c, Supplementary Fig. 3b).

[1]Department of Geology and Geophysics, Yale University, PO Box 208109, New Haven, Connecticut 06520, USA. [2]Research Unit Palaeontology, Department of Geology and Soil Science, Ghent University, Krijgslaan 281/S8, B-9000 Ghent, Belgium. [3]Yale Peabody Museum of Natural History, Yale University, New Haven, Connecticut 06520, USA.

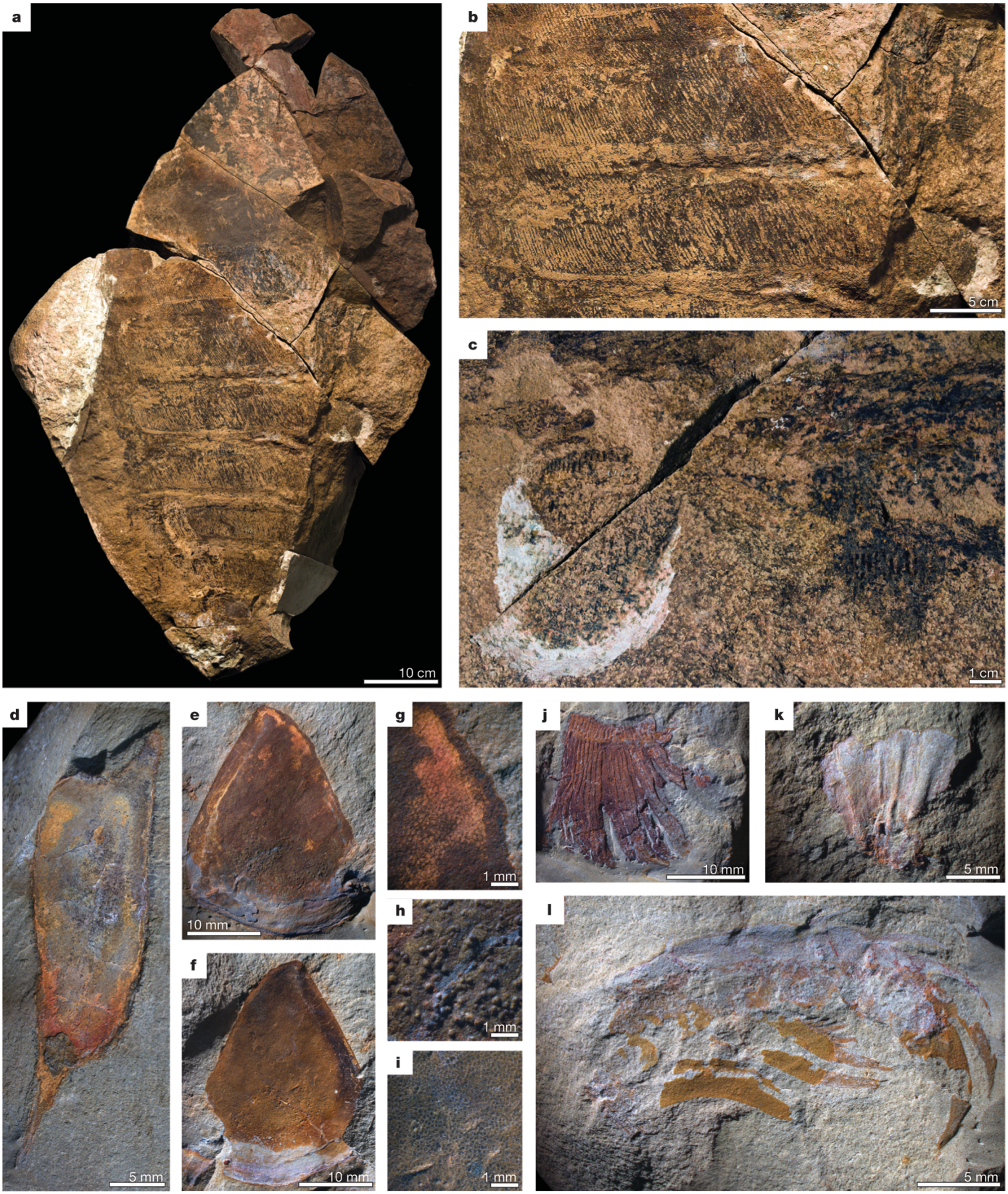


Figure 1 | Anomalocaridid specimens from the Lower Ordovician Fezouata formations. Drawings by *camera lucida* of the specimens are provided in Supplementary Fig. 4. **a–c**, Giant, near-complete specimen preserved in a concretion, dorsal side exposed, Excavation 1 (YPM 226437). **a**, Entire specimen. **b**, Segments 5 and 6 showing blades. **c**, Right lobes 5 and 6 showing rays. **d**, Left lateral 'P-element' of carapace, part, Excavation 2 (YPM 227518). **e–i**, Central 'H-element' of carapace, Excavation 2 (YPM 227517). **e**, Part showing evidence of a healed injury, posterior right. **f**, Counterpart. **g**, Reticulate structure on the anterior right of the part. **h**, Coarse tubercles on the mid-posterior of the part. **i**, Fine tubercles on the counterpart. **j**, Fragment of gill, part, Excavation 2 (YPM 227934). **k**, Incomplete oral circlet, part, Excavation 3 (YPM 227643). **l**, Great appendage, part, Excavation 5 (YPM 227644).

Smaller individuals were represented in the Fezouata Formation at other excavations (Supplementary Fig. 1). Excavation 2 yielded an elongate valve-like sclerite (41 mm long including spines), with a robust triangular projection at one end and a long slender pointed projection at the other (Fig. 1d, Supplementary Fig. 4c). A sub-triangular sclerite (Fig. 1e–i, Supplementary Fig. 4b, c), 32 mm long, was found at the same excavation. It preserves a reticulate structure on the part (Fig. 1e, g) but not on the counterpart (Fig. 1f): this may be an internal feature of the cuticle revealed by exfoliation. Coarse tubercles are scattered in the posterior half (Fig. 1e, h) and fine tubercles are closely spaced elsewhere (Fig. 1i).

Excavation 2 also yielded a relatively broad jointed rachis about 15 mm long, to which long, flat, blunt-tipped, overlapping flexible blades are attached, one on each podomere (Fig. 1j, Supplementary Fig. 4d). These structures are similar to those on the segments of the largest specimens (Fig. 1a, b; Supplementary Figs 2, 3a). Two less well-preserved fragments of similar morphology were also recovered from the slightly younger Excavation 4.

An incomplete oral circlet consisting of one major and two minor plates (Fig. 1k, Supplementary Fig. 4e) was found at Excavation 3. The major plate (preserved length ~13 mm) shows two narrow radial strengthening ridges; only one wider ridge is present on the minor plates. The teeth were broken away as the shale was split to reveal the fossil. Extrapolation suggests a diameter of the complete circlet of slightly more than 30 mm.

Two specimens of a great appendage were also recovered. The number of podomeres and the morphology of the spines indicates that they belong to an anomalocaridid. The larger specimen from Excavation 5 (preserved length ~34 mm along the curve) is incomplete at both ends, but preserves evidence of 9 podomeres (Fig. 1l, Supplementary Fig. 4f). A distal spine on the dorsal side is small and more dorsally directed on the more proximal podomeres; these spines become more robust and near parallel to the appendage on successively distal podomeres. A very robust, slightly recurving, long spine projects ventrally from the mid-length of the more proximal podomeres but may have been absent from the distalmost four. This ventral spine terminates in a curved point, and bears five or six secondary spines along its proximal length. A fine bundle of spines or setae is associated with the proximal part of this larger specimen but it is not known whether this is simply a chance association. A smaller specimen from Excavation 3 (preserved length ~16 mm along the curve) comprises 10 or 11 podomeres (Supplementary Fig. 3c, d); ventral spines appear to be present only on the median 4 podomeres.

In summary, the large articulated bodies all derive from Excavation 1, whereas the small carapace elements were found at an adjacent site of similar age (Excavation 2) together with a fragment of rachis with blades. Blade-like structures were collected from the younger Excavation 4 whereas the similar aged Excavations 3 and 5 yielded raptorial appendages, associated with a partial oral circlet at Excavation 3. It is unclear whether the Moroccan specimens represent more than one type of anomalocaridid: a formal description awaits the discovery of more material.

The blade-like structures on the surface of the segments and in the fragmentary specimens are similar to those in *Hurdia* and *Laggania*[6], and may be comparable to the blades in *Opabinia*[27]. A transverse arrangement of blades across the trunk similar to that in the large Moroccan specimens has been described and reconstructed on the dorsal surface of *Laggania*[14,25,26]. The attachment of the blades to a rachis is similar to the arrangement in *Hurdia*, and the transverse mineralized structures in *Laggania* may be homologous. The lateral lobes show rays, as observed in *Laggania*. The carapace elements are strikingly similar to the central 'H-element' and lateral 'P-element' that make up the tripartite *Hurdia* headshield[6]. Although they differ in outline from those typical of *Hurdia victoria* from the Burgess Shale, the H-element preserves an internal reticulate structure similar to that in *Hurdia* sclerites[6]. The great appendages resemble morph B in *Hurdia*[6] although they differ in detail.

The specimens from Excavation 1 represent the largest articulated anomalocaridid specimens known, and even though the size of the head is unknown, it is likely that the largest Moroccan individual was significantly larger than its Burgess Shale counterparts[3]. These anomalocaridids are rivalled in size among arthropods only by pterygotid eurypterids[28] and the terrestrial arthropleurids[29] (although extrapolation based on remains of the oral circlet from the early Cambrian Chengjiang biota of China[1] has been used to infer an even larger anomalocaridid, these estimates are uncertain[14]). They dwarf the other organisms known from the Fezouata Biota: the largest trilobites are nileids, asaphids and dikelokephalinids, which reach a maximum of ~30 cm (ref. 24), less than 30% of the length of the anomalocaridids.

Like other anomalocaridids, the Moroccan examples are assumed to have been swimming predators. The great appendages seem to be adapted to entrap prey and to help to transfer it to the mouth. No biomineralized organisms with injuries compatible with the anomalocaridid oral circlet have been recovered from the Fezouata formations, but it is likely that the Moroccan anomalocaridids, like their Cambrian predecessors, fed mostly on unmineralized organisms[3]. The size and associated food requirements of the largest Fezouata individuals might imply low population densities. The association of at least five individuals in close proximity at Excavation 1 may indicate an abundant food source, or congregation for some other purpose, such as moulting or mating.

The most striking feature of the Moroccan specimens, apart from their size, are the series of dorsal blades. These specimens provide the most compelling evidence yet that such structures traversed the entire trunk of some anomalocaridid taxa[19,25,26]. They do not resemble scales[25] and their interpretation as gills is more plausible, even if their unprotected position is anomalous. There is no sign of any separation of the left and right halves as in *Hurdia*[6]. While the blades may originate as exites[6], the rachis to which they connect appears to be attached across the dorsal width of the trunk in the middle of the intersegment areas in a manner unknown in euarthropods; it is unclear whether the rachis also attached to the base of the lobes. If the function of the blades was respiratory, they may reflect low oxygen conditions or the high oxygen requirement of a large active swimmer; their dorsal position would have assured continued oxygenation when in close proximity to anoxic bottom conditions.

The Fezouata discoveries extend the range of unequivocal anomalocaridids by about 30 million years ago, from the middle Cambrian to the Early Ordovician, and provide the only temporal link between the Cambrian occurrences and the Early Devonian great appendage arthropod *Schinderhannes*, which retains some anomalocaridid characters[5]. The Moroccan occurrences show that anomalocaridids were the largest organisms in some ecosystems even in the Ordovician, and were presumably at the top of the food chain. With the exception of the singular Polish and possible Czech occurrences, which are from intermediate and polar palaeolatitudes respectively, all Cambrian anomalocaridid localities are situated in the palaeotropics. The Fezouata formations, in contrast, were deposited at a high polar southern palaeolatitude, confirming a global distribution of anomalocaridids, at least latitudinally, during the early Palaeozoic, as observed for many other taxa of Burgess Shale type[8]. The demise of anomalocaridids may have been associated with the diversification of large predatory eurypterids and stem cephalopods during the Great Ordovician Biodiversification Event[30].

METHODS SUMMARY

All figured specimens are housed in the collections of the Yale Peabody Museum of Natural History. Locality details are kept at the Museum, and can be provided on request. Small specimens were prepared with scalpels and fine needles under high magnification using Leica MZ6 and MZ16 stereomicroscopes and, when necessary, repaired using cyanoacrylate glue. The large specimens were prepared with the aid of fine chisels and electric engravers, and assembled using cyanoacrylate glue and epoxy. Interpretative drawings were made with a *camera lucida* attached to a Leica MZ6 stereomicroscope. Photographs of the large specimens were made with a

Canon EOS 300D digital reflex camera with a Sigma EX 50 mm f 2.8 DG macro lens stopped down to f5 for maximum sharpness, while smaller specimens were imaged using a Leica MZ16 stereomicroscope with Leica DFC 425 digital camera. All photographs were taken with crossed polarizers, and, in addition, the specimen in Supplementary Fig. 3c was photographed under ethanol. Digital photographs were processed in Adobe Photoshop CS2 and CS3. To maximize depth of field, between 10 and 30 images were stacked using CombineZP and Helicon Focus Pro software. Figures 1a, b, d, e, f, j, l and Supplementary Figs 2a, c, and 3a are composite images, stitched together using Adobe Photoshop CS3 and Microsoft ICE. Elemental composition was analysed in a Philips XL 30 environmental scanning electron microscope (ESEM) equipped with an energy dispersive X-ray analyser (EDX).

Received 19 January; accepted 9 February 2011.

1. Chen, J. Y., Ramsköld, L. & Zhou, G. Q. Evidence for monophyly and arthropod affinity of Cambrian giant predators. *Science* **264,** 1304–1308 (1994).
2. Collins, D. The ''evolution'' of *Anomalocaris* and its classification in the arthropod Class Dinocarida (nov.) and Order Radiodonta (nov.). *J. Paleontol.* **70,** 280–293 (1996).
3. Whittington, H. B. & Briggs, D. E. G. The largest Cambrian animal, *Anomalocaris,* Burgess Shale, British Columbia. *Phil. Trans. R. Soc. Lond. B* **309,** 569–609 (1985).
4. Budd, G. E. A Cambrian gilled lobopod from Greenland. *Nature* **364,** 709–711 (1993).
5. Kühl, G., Briggs, D. E. G. & Rust, J. A great appendage arthropod with a radial mouth from the Lower Devonian Hunsrück Slate, Germany. *Science* **323,** 771–773 (2009).
6. Daley, A. C., Budd, G. E., Caron, J.-B., Edgecombe, G. D. & Collins, D. The Burgess Shale anomalocaridid *Hurdia* and its significance for early euarthropod evolution. *Science* **323,** 1597–1600 (2009).
7. Briggs, D. E. G. & Robison, R. A. Exceptionally preserved nontrilobite arthropods and *Anomalocaris* from the Middle Cambrian of Utah. *Univ. Kansas Paleont. Contrib.* **111,** 1–23 (1984).
8. Van Roy, P. *et al.* Ordovician faunas of Burgess Shale type. *Nature* **465,** 215–218 (2010).
9. Whiteaves, J. F. Description of a new genus and species of phyllocarid crustacean from the Middle Cambrian of Mount Stephen, British Columbia. *Can. Rec. Sci.* **5,** 205–208 (1892).
10. Walcott, C. D. Middle Cambrian Merostomata. Cambrian geology and paleontology II. *Smithson. Misc. Coll.* **57,** 17–40 (1911).
11. Walcott, C. D. Middle Cambrian Holothurians and Medusae. Cambrian geology and paleontology II. *Smithson. Misc. Coll.* **57,** 41–68 (1911).
12. Walcott, C. D. Middle Cambrian Branchiopoda, Malacostraca, Trilobita, and Merostomata. Cambrian geology and paleontology II. *Smithson. Misc. Coll.* **57,** 145–228 (1912).
13. Rolfe, W. D. I. Two new arthropod carapaces from the Burgess Shale (Middle Cambrian) of Canada. *Breviora* **160,** 1–9 (1962).
14. Hou, X., Bergström, J. & Yang, J. Distinguishing anomalocaridids from arthropods and priapulids. *Geol. J.* **41,** 259–269 (2006).
15. Chen, J., Waloszek, D. & Maas, A. A new 'great-appendage' arthropod from the Lower Cambrian of China and homology of chelicerate chelicerae and raptorial antero-ventral appendages. *Lethaia* **37,** 3–20 (2004).
16. Dzik, J. & Lendzion, K. The oldest arthropods of the East European platform. *Lethaia* **21,** 29–38 (1988).
17. Daley, A. C. & Budd, G. E. New anomalocaridid appendages from the Burgess Shale, Canada. *Palaeontology* **53,** 721–738 (2010).
18. Ponomarenko, A. G. First record of Dinocarida from Russia. *Paleontol. J.* **44,** 503–504 (2010).
19. Hou, X., Bergström, J. & Ahlberg, P. *Anomalocaris* and other large animals in the Lower Cambrian Chengjiang fauna of southwest China. *Geol. För. Förh.* **117,** 163–183 (1995).
20. Daley, A. C. & Peel, J. S. A possible anomalocaridid from the Cambrian Sirius Passet Lagerstätte, north Greenland. *J. Paleontol.* **84,** 352–355 (2010).
21. Masiak, M. & Zylinska, A. Burgess Shale-type fossils in Cambrian sandstones of the Holy Cross Mountains. *Acta Palaeontol. Pol.* **39,** 329–340 (1994).
22. Van Roy, P. & Tetlie, O. E. A spinose appendage fragment of a problematic arthropod from the Early Ordovician of Morocco. *Acta Palaeontol. Pol.* **51,** 239–246 (2006).
23. Vinther, J., Van Roy, P. & Briggs, D. E. G. Machaeridians are Palaeozoic armoured annelids. *Nature* **451,** 185–188 (2008).
24. Fortey, R. A. A new giant asaphid trilobite from the Lower Ordovician of Morocco. *Mem. Assoc. Austral. Palaeontol.* **37,** 9–16 (2009).
25. Bergström, J. *Opabinia* and *Anomalocaris,* unique Cambrian 'arthropods'. *Lethaia* **19,** 241–246 (1986).
26. Bergström, J. The Cambrian *Opabinia* and *Anomalocaris. Lethaia* **20,** 187–188 (1987).
27. Budd, G. E. & Daley, A. C. The lobes and lobopods of *Opabinia regalis* of the middle Cambrian Burgess Shale. *Lethaia* (in the press).
28. Braddy, S. J., Poschmann, M. & Tetlie, O. E. Giant claw reveals the largest ever arthropod. *Biol. Lett.* **4,** 106–109 (2008).
29. Hahn, G., Hahn, R. A. & Brauckmann, C. Zur Kenntis von *Arthropleura* (Myriapoda; Ober Karbon). *Geol. Paleontol.* **20,** 125–137 (1986).
30. Servais, T., Owen, A. W., Harper, D. A. T., Kröger, B. & Munnecke, A. The Great Ordovician Biodiversification Event (GOBE): the palaeoecological dimension. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **294,** 99–119 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements S. Butts (Yale Peabody Museum of Natural History) provided access to specimens. M. Ben Said Ben Moula originally discovered the specimens and made them available for study; he, together with J. P. Botting-Muir and L. A. Botting-Muir, P. J. Orr, C. Upton and J. Vinther also assisted with fieldwork, and B. Tahiri arranged logistical support. J. W. Hagadorn shared information on anomalocaridid size and R. R. Gaines discussed aspects of the taphonomy. E. Champion helped with the preparation of figures. This research was supported by a National Geographic Society Research and Exploration grant and by Yale University.

Author Contributions The authors contributed equally to interpreting the fossils and writing the paper. P.V.R. played the primary role in field work, and prepared, photographed and drew the specimens.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.E.G.B. (derek.briggs@yale.edu).

LETTER

doi:10.1038/nature09946

Learning-related feedforward inhibitory connectivity growth required for memory precision

Sarah Ruediger[1]*, Claudia Vittori[1,2]*, Ewa Bednarek[1], Christel Genoud[1], Piergiorgio Strata[2], Benedetto Sacchetti[2] & Pico Caroni[1]

In the adult brain, new synapses are formed and pre-existing ones are lost, but the function of this structural plasticity has remained unclear[1–5]. Learning of new skills is correlated with formation of new synapses[6–8]. These may directly encode new memories, but they may also have more general roles in memory encoding and retrieval processes[2]. Here we investigated how mossy fibre terminal complexes at the entry of hippocampal and cerebellar circuits rearrange upon learning in mice, and what is the functional role of the rearrangements. We show that one-trial and incremental learning lead to robust, circuit-specific, long-lasting and reversible increases in the numbers of filopodial synapses onto fast-spiking interneurons that trigger feedforward inhibition. The increase in feedforward inhibition connectivity involved a majority of the presynaptic terminals, restricted the numbers of c-Fos-expressing postsynaptic neurons at memory retrieval, and correlated temporally with the quality of the memory. We then show that for contextual fear conditioning and Morris water maze learning, increased feedforward inhibition connectivity by hippocampal mossy fibres has a critical role for the precision of the memory and the learned behaviour. In the absence of mossy fibre long-term potentiation in *Rab3a*$^{-/-}$ mice[9], c-Fos ensemble reorganization and feedforward inhibition growth were both absent in CA3 upon learning, and the memory was imprecise. By contrast, in the absence of adducin 2 (Add2; also known as β-adducin)[10] c-Fos reorganization was normal, but feedforward inhibition growth was abolished. In parallel, c-Fos ensembles in CA3 were greatly enlarged, and the memory was imprecise. Feedforward inhibition growth and memory precision were both rescued by re-expression of Add2 specifically in hippocampal mossy fibres. These results establish a causal relationship between learning-related increases in the numbers of defined synapses and the precision of learning and memory in the adult. The results further relate plasticity and feedforward inhibition growth at hippocampal mossy fibres to the precision of hippocampus-dependent memories.

To determine whether hippocampus-dependent learning[11–13] may produce structural rearrangements in hippocampal large mossy fibre terminal (LMT) components involved in feedforward excitation and/or feedforward inhibition in CA3 (ref. 14) (Fig. 1a and Supplementary Material), we analysed GFP-positive LMTs in the dorsal hippocampus of *Thy1-mGFP(Lsi1)* reporter mice[5] that had been subjected to contextual fear conditioning, a one-trial learning protocol (Methods). Fear conditioning led to a robust increase in the average number of filopodia per LMT (1.82-fold, $P < 0.001$; feedforward inhibition connectivity; Fig. 1b, c and Supplementary Fig. 2a), and to a less pronounced increase in the average numbers of Bassoon-positive putative release sites per core LMT[15] (1.31-fold, $P < 0.01$; feedforward excitation connectivity; Supplementary Fig. 2a). By contrast, there was no change in the densities of LMTs in CA3b at any time upon fear conditioning (Supplementary Fig. 2a). The filopodia contacted spine-free dendrites of parvalbumin-positive interneurons in CA3b (Fig. 1d, e and Supplementary Fig. 3a), indicating that they induce feedforward inhibition through fast-spiking interneurons[16–18]. To estimate the fraction of LMTs in CA3b with altered contents of filopodia, we analysed LMT/filopodia distributions in naive, control and fear-conditioned mice. Shifts in the fractions of LMTs with no filopodia and with more than four filopodia revealed that, on average, at least 45% of the LMTs established increased numbers of filopodia as a consequence of fear conditioning (Fig. 1f).

To determine whether an increase in stratum lucidum feedforward inhibition connectivity may be generally associated with hippocampal learning, we analysed mice that underwent a Morris water maze incremental learning protocol. Filopodial contents were only slightly increased over naive values during the first 3–4 days of training, whereas they increased markedly between days 4 and 8 (Fig. 1g). Again, we detected no changes in the densities of LMTs in CA3b upon Morris water maze learning (not shown). Testing mice for the memory of the platform position revealed that this reference memory only began to differ from chance after 3 days of training (Fig. 1h). The reference memory reached plateau values at day 8 (Fig. 1h), suggesting that filopodial growth correlated with the establishment of a precise spatial memory in the Morris water maze test. The reference memory of the platform position persisted for at least 45 days after cessation of the training and, unlike in the fear conditioning experiment, raised filopodia per LMT values also persisted for at least 45 days (Fig. 1h; day 53 values). As in the fear conditioning experiment, a large fraction of the LMTs exhibited higher filopodial contents at plateau values (Fig. 1i).

To determine whether learning-related induction of feedforward inhibition connectivity growth might be a general phenomenon not restricted to spatial learning in the hippocampus, we analysed mossy fibre terminals in the cerebellar cortex, which also consist of powerful large core structures associated with filopodia[19]. Cued fear conditioning, in which animals learn that a tone predicts an aversive stimulus, involves plasticity in cerebellar cortex lobule 5, but not lobule 9 (ref. 20). In parallel, cued fear conditioning led to a robust and reversible increase of filopodial numbers per mossy fibre terminal in lobule 5, but not lobule 9 (Fig. 2a, d). In a second set of experiments, we trained mice to balance on an accelerating rotating rod (rotarod). This cerebellum-dependent motor skill task involved incremental learning over 4–6 days, which was accompanied by a parallel increase in the filopodial contents of mossy fibre terminals in lobule 9, but not lobule 5 (Fig. 2b, d). At least for the Golgi cells that could be visualized with the marker RC3, mossy fibre terminal filopodia extended along their dendrites, and established numerous varicosities, where synaptic markers co-distributed (Fig. 2c and Supplementary Fig. 4). More than 95% of the filopodial varicosities within a granule cell layer volume exhibiting an RC3-positive Golgi cell made putative synaptic contacts with that Golgi cell. Therefore, learning is specifically correlated with the growth of feedforward inhibition connectivity in both hippocampal and cerebellar circuits.

We next sought to determine what might be the function of the learning-related growth in feedforward inhibition connectivity. In the fear conditioning experiments, the excess filopodia were lost within 8–10 days after learning, and filopodial retention was prolonged upon re-exposure to context leading to extinction (Fig. 3a), indicating that

[1]Friedrich Miescher Institute, Maulbeerstrasse 66, CH-4058 Basel, Switzerland. [2]Department of Neuroscience and National Institute of Neuroscience-Italy, C.so Raffaello 30, 10125 Torino, Italy.
*These authors contributed equally to this work.

the excess filopodia are not a requirement for expression of the fear memory. Testing of individual mice during the Morris water maze training protocol revealed a strong correlation between the reference memory of the platform position and mean filopodial contents per LMT for individual mice (Fig. 3b), indicating that the extent of filopodial growth was correlated to the precision of the learning. We therefore monitored generalization, that is, decreased behavioural precision of the fear memory in the contextual fear conditioning experiment. In agreement with previous reports[21,22], generalization of the memory for context in fear conditioning was not detectable during the first 5–7 days after learning, but was detected at longer intervals after fear conditioning as an enhanced freezing response and reduced exploratory activity in a neutral context (Fig. 3c). A brief re-exposure of mice to training context in the absence of the aversive stimulus at 15 days after learning produced a suppression of generalization at retest, which lasted 8–12 days (Fig. 3d). In parallel, training context re-exposure induced a pronounced re-induction of the filopodial response, which again lasted for 7–10 days (Fig. 3d). By contrast, exposure to a neutral context affected neither generalization nor filopodial growth (Fig. 3d), suggesting that retrieval of the specific memory was necessary to re-induce feedforward inhibition connectivity growth in hippocampal CA3, and to suppress generalization.

To investigate a possible functional correlate of feedforward inhibition connectivity growth, we analysed c-Fos-positive pyramidal neurons in CA3b in the contextual fear conditioning experiment[23]. On day 0, mice were exposed to the training context without or with aversive stimulus. In the absence of aversive conditioning, re-exposure on day 1 to either the training context or a neutral context produced closely comparable increases in the fractions of pyramidal neurons with high and intermediate c-Fos signals when compared to naive cage control mice (Fig. 4a). In stark contrast, association of the training context with an aversive stimulus led to a specific and robust relative increase in the number of pyramidal neurons expressing high c-Fos signals upon recall of the memory in the training context, and to a marked reduction of the high and medium c-Fos signals upon exposure to the neutral context (Fig. 4a). Recall in the training context at day 15 led to decreased high-signal c-Fos neurons, whereas exposure to a neutral context at day 15 led to markedly increased low-signal c-Fos neurons (Fig. 4b). Notably, in parallel to increased filopodial numbers and the re-establishment of memory precision, memory recall in the training context at day 15 after fear conditioning suppressed excess responses upon subsequent exposure to a neutral context (Fig. 4b).

To address the role of mossy fibres and their plasticity in fear memory precision, we carried out fear conditioning experiments in *Rab3a*$^{-/-}$ mice, which specifically lack long-term potentiation (LTP) at mossy fibres, but not at other synapses in the hippocampus[9]. We found that in the absence of Rab3a, mice learned the relationship between the training context and the aversive stimulus, but already generalized 1 day after fear conditioning (Fig. 4c). In parallel, *Rab3a*$^{-/-}$ mice lacked any learning-related increase in putative release sites at core LMTs, or any learning-related increase in filopodia numbers at LMTs in CA3 (Fig. 4c). Furthermore, analysis of c-Fos-positive neurons upon recall 1 day after learning revealed a complete absence of ensemble activity rearrangements in CA3 upon fear conditioning, leading to comparable contents of c-Fos-positive neurons upon re-exposure to the training context or exposure to an unrelated neutral context, regardless of associative learning through aversive pairing (Fig. 4d). These results indicate that synaptic plasticity at LMTs in CA3 is required to re-organize pyramidal neuron ensemble activity in CA3 upon fear conditioning learning, to establish a precise memory of context in the hippocampus, and to induce learning-related feedforward inhibition growth.

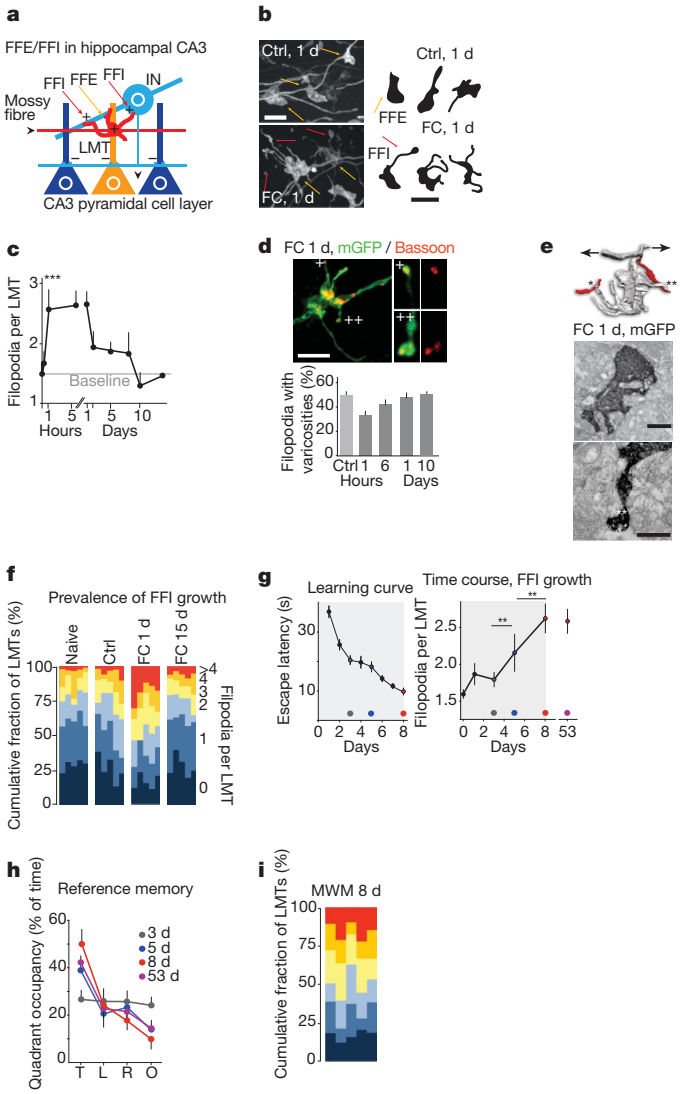
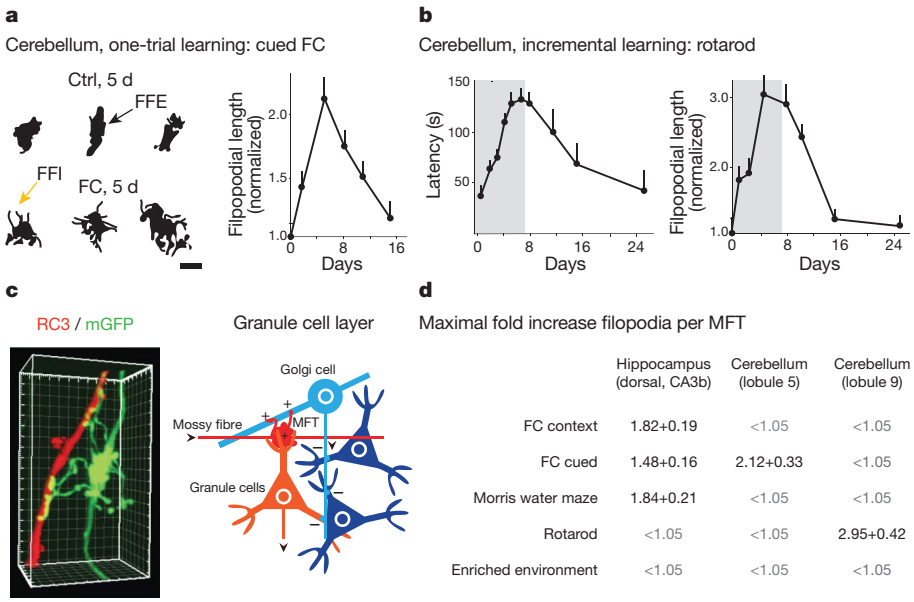


Figure 1 | Learning-related feedforward inhibition connectivity growth in the hippocampus. **a**, Schematic of hippocampal feedforward excitation (FFE) and feedforward inhibition (FFI) circuit in stratum lucidum of CA3. IN, inhibitory interneuron. **b**–**f**, Feedforward inhibition growth at hippocampal mossy fibre LMTs upon contextual fear conditioning. **b**, Micrographs and representative camera lucidas of mGFP-labelled mossy fibres and LMTs in hippocampal stratum lucidum (CA3b). Yellow arrows, core LMTs; red arrows, filopodia. Ctrl; control; FC, fear conditioning. **c**, Average filopodia/LMT values upon fear conditioning. $N = 5$ mice (100 LMTs each). $^{***}P < 0.001$. **d**, Filopodial synapses upon fear conditioning. Overview panel shows maximal intensity projection of mGFP-positive LMT with four filopodia. Detail panels show single confocal planes of two of the filopodia (+ and ++); Bassoon channel masked using three-dimensional isosurface of GFP-positive LMT. Bar diagram shows fraction of LMT filopodia with varicosities as a function of time upon fear conditioning ($N = 3$, 100 LMTs). **e**, Filopodia upon fear conditioning learning contact spine-free dendrites. Immuno-electron microscopy of mGFP-positive LMT with four filopodia, 1 day after fear conditioning. Top, three-dimensional reconstruction of immuno-labelled LMT (red, spine-free dendrites contacted by two of the filopodia in the example (marked by one and two asterisks, respectively). Centre, immuno-labelled LMT. Bottom, filopodium with contact is marked by two asterisks. **f**, Distributions of filopodia per LMT contents for individual mice. $N = 100$ LMTs. Relative contents of LMTs with 0, 1, 2, 3, 4, >4 filopodia as a fraction of the total LMT population. Vertical rows, individual mice. **g**–**i**, Feedforward inhition growth at hippocampal mossy fibre LMTs upon Morris water maze (MWM) training. **g**, Learning curve and time course of feedforward inhibition growth. $N = 5$ mice (100 LMTs each). Grey area shows daily training period. The circles highlight the positions on the curves as compared to reference memory (right). **h**, Reference memory at 3, 5, 8 and 53 days. Percentage of time spent by the mice in target (T), left (L), right (C) and opposite (O) quadrants. $N = 5$ mice. **i**, Filopodia per LMT distributions after 8 days of training, as described in **b**. Scale bars, 5 μm (**b**, **d**, top and **g**), 1 μm (**d**, bottom centre) and 0.5 μm (**d**, bottom right).



	Hippocampus (dorsal, CA3b)	Cerebellum (lobule 5)	Cerebellum (lobule 9)
FC context	1.82+0.19	<1.05	<1.05
FC cued	1.48+0.16	2.12+0.33	<1.05
Morris water maze	1.84+0.21	<1.05	<1.05
Rotarod	<1.05	<1.05	2.95+0.42
Enriched environment	<1.05	<1.05	<1.05

Figure 2 | Specificity of learning-related feedforward inhibition growth. **a**, **b**, Learning-related feedforward inhibition connectivity growth in cerebellar cortex. **a**, Feedforward inhibition growth at lobule 5 cerebellar cortex mossy fibre terminals upon cued fear conditioning. Labelling as in Fig. 1b. Scale bar, 10 µm. **b**, Feedforward inhibition growth at lobule 9 cerebellar cortex mossy fibre terminals (MFTs) upon rotarod learning. Labelling as in Fig. 1c. **c**, In cerebellar cortex, mossy fibre terminal filopodia contact inhibitory Golgi cells. Left, three-dimensional rendering of contacts by mossy fibre terminal filopodia onto RC3-positive Golgi cell dendrite. Right, feedforward excitation/feedforward inhibition circuit in granule cell layer of cerebellar cortex. **d**, Specific relationship between learning and feedforward inhibition growth. Average fold increase values at peak response (fear conditioning hippocampus, 1 day; fear conditioning cerebellum, 2 days; Morris water maze, 8 days; rotarod, 5 days). $N = 5$, 100 LMTs or mossy fibre terminals each.

To test the notion that learning-related feedforward inhibition growth is necessary for memory precision, we then carried out learning experiments in *Add2* knockout mice[10], which exhibit early LTP, but have a defect in synapse stabilization due to impaired linkage between the cell membrane cortex and the actin cytoskeleton[24] In naive $Add2^{-/-}$ mice, average values of filopodia per LMT were closely comparable to those in wild-type mice. Unlike $Rab3a^{-/-}$ mice, $Add2^{-/-}$ mice did exhibit enhanced putative release sites per core LMT upon fear conditioning (Supplementary Material), but they completely failed to establish higher numbers of filopodia upon fear conditioning (Fig. 5a). In parallel, and like $Rab3a^{-/-}$ mice, $Add2^{-/-}$ mice learned to associate fear with context, but the memory was imprecise and mice already generalized 1 day after fear conditioning (Fig. 5a). Comparable findings were obtained for Morris water maze and rotarod learning in $Add2^{-/-}$ mice (Fig. 5b and Supplementary Material). Absence of learning-related feedforward inhibition connectivity growth in $Add2^{-/-}$ mice is thus correlated with poor precision of the learned memory in the fear conditioning and Morris water maze paradigms, and with a near to complete failure to learn the rotarod task.

We then investigated c-Fos-positive CA3 pyramidal neuron ensembles in response to fear conditioning in the $Add2^{-/-}$ mice. In stark contrast to $Rab3a^{-/-}$ mice lacking mossy fibre LTP, and consistent with increased feedforward excitation connectivity, $Add2^{-/-}$ mice exhibited c-Fos ensemble reorganization responses in CA3 that were qualitatively closely comparable to those in wild-type mice (Fig. 5c). Remarkably, however, net total numbers of c-Fos-positive neurons were more than 2.5 times higher for each experimental condition in $Add2^{-/-}$ mice compared to wild-type mice (Fig. 5c). By contrast, numbers of c-Fos-positive pyramidal neurons in naive $Add2^{-/-}$ mice were not higher than those in naive wild-type mice, indicating that the mutant mice did not just exhibit raised levels of c-Fos in CA3 neurons (Fig. 5c).

In wild-type mice, a reorganization of training context/neutral context ensembles upon fear conditioning was also detected in granule cells, but it was much less marked than in CA3 (Fig. 5d). Notably, however, and in stark contrast to CA3, distributions and numbers of c-Fos positive granule cells in $Add2^{-/-}$ mice were not different from those in wild-type mice for all experimental conditions tested (Fig. 5d). Therefore, $Add2^{-/-}$ mice re-organized their CA3 pyramidal neuron ensembles like wild-type mice, but failed to restrict the numbers of

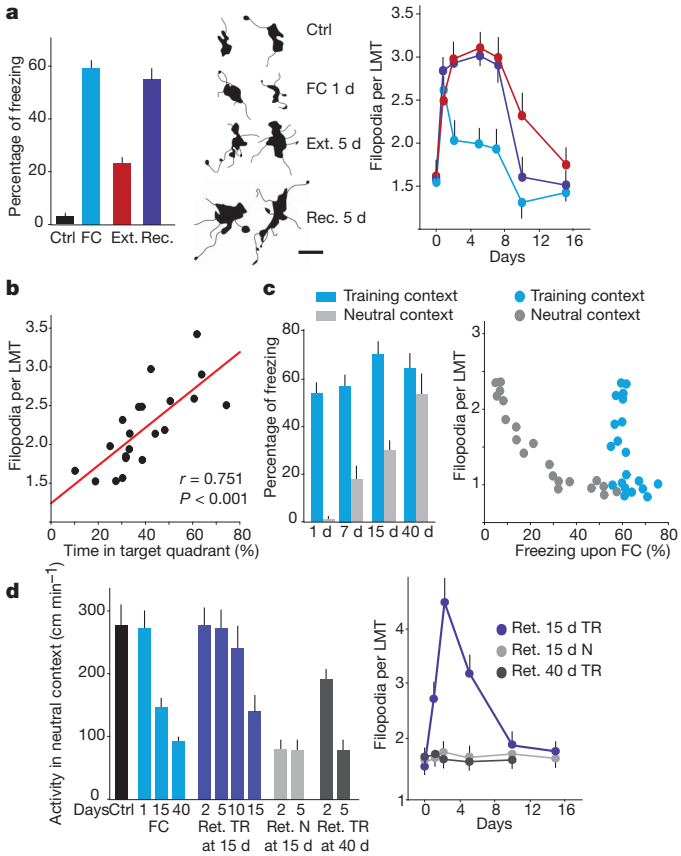


Figure 3 | Correlation between feedforward inhibition growth and quality of hippocampal learning and memory. **a**, Memory retrieval prolongs peak levels of feedforward inhibition growth upon cued fear conditioning (FC). Pale blue: fear conditioning, no recall (at 1 day); red: fear conditioning followed by extinction (Ext.) at 5 h and 24 h (at 5 days); violet: fear conditioning followed by recall (Rec.) at 5 h and 24 h (at 5 days). $N = 5$ mice (100 LMTs each). Scale bar, 5 µm. **b**, Correlation between reference memory precision and average filopodial contents per LMT in Morris water maze task. Dots show individual mice analysed between day 1 and day 8 of the training procedure (100 LMTs each). **c**, Time-dependent generalization upon contextual fear conditioning learning. Right, dots represent average values for individual mice at different times after fear conditioning learning (100 LMTs each). **d**, Re-growth of filopodia and re-contextualization upon retrieval of training context memory (Ret. TR) versus retrieval of neutral context (Ret. N). Left, exploratory activity in neutral context as a function of days after last manipulation. Error bars show mean ± s.e.m.

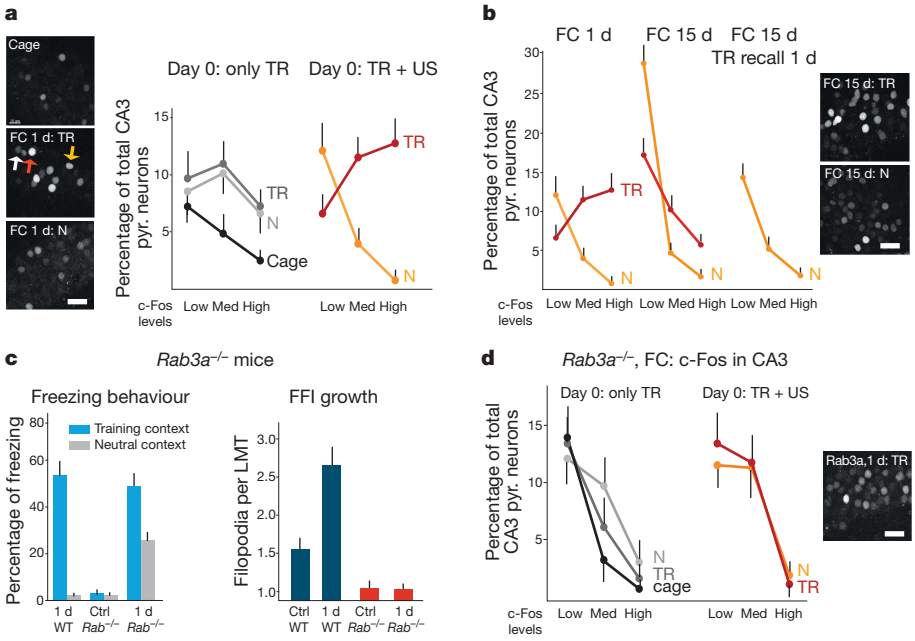


Figure 4 | Relationship between induction of c-Fos in CA3 pyramidal neurons and behavioural memory precision upon contextual fear conditioning. **a**, c-Fos immunoreactivity in CA3 pyramidal neurons upon exposure to training context (TR) or neutral context (N), with or without aversive association. Panels show representative examples of c-Fos immunoreactivity in CA3b. c-Fos neurons classified as weak (white arrow), medium (yellow arrow), strong (red arrow). $N = 3$, 500 pyramidal (pyr.) neurons each. US, unconditioned, aversive stimulus. **b**, c-Fos immunoreactivity in CA3 pyramidal neurons 15 days after fear conditioning: effect of recall with training context. Details as in **a**. $N = 1{,}000$–$1{,}500$ pyramidal neurons, from 3 mice each. **c**, Generalization and absence of learning-induced feedforward inhibition growth in *Rab3a*$^{-/-}$ mice. $N = 5$ mice (100 LMTs each). **d**, c-Fos immunoreactivity in CA3 pyramidal neurons of *Rab3a*$^{-/-}$ mice upon fear conditioning. Details as in **a**. $N = 1{,}000$–$1{,}500$ pyramidal neurons, from 3 mice each. Scale bars, 20 μm. Error bars show mean ± s.e.m.

activated pyramidal neurons in CA3 upon stimuli, which is consistent with a complete absence of feedforward inhibition connectivity growth at LMTs. Furthermore, c-Fos activation patterns in CA3 correlated with memory precision, whereas those in dentate gyrus did not, suggesting that the absence of Add2 in mossy fibres and their LMTs may account for the impaired memory precision in *Add2*$^{-/-}$ mice.

To establish a causal link between learning-related feedforward inhibition growth at LMTs and memory precision, we determined whether re-expression of Add2 specifically in granule cells and their mossy fibres was sufficient to rescue filopodial growth and memory precision upon fear conditioning. To achieve specific re-expression in the adult, we expressed Add2 selectively in the dentate gyrus[15] of *Add2*$^{-/-}$ mice using a lentiviral construct. One month after viral transduction, 15–22% of granule cells throughout the entire hippocampus exhibited virus-driven gene expression, whereas expression outside the dentate gyrus was extremely rare (Fig. 5e). The re-introduction of Add2 in mossy fibres was sufficient to rescue filopodial growth at LMTs of transduced granule cells in response to fear conditioning (Fig. 5f). Most notably, and in parallel to restored feedforward inhibition growth, re-expression of Add2 in granule cells rescued behavioural contextualization upon fear conditioning (Fig. 5g).

Our results establish a causal relationship between learning-associated structural alterations in identified circuit connectivity and a specific behavioural output. We provide evidence that increased feedforward

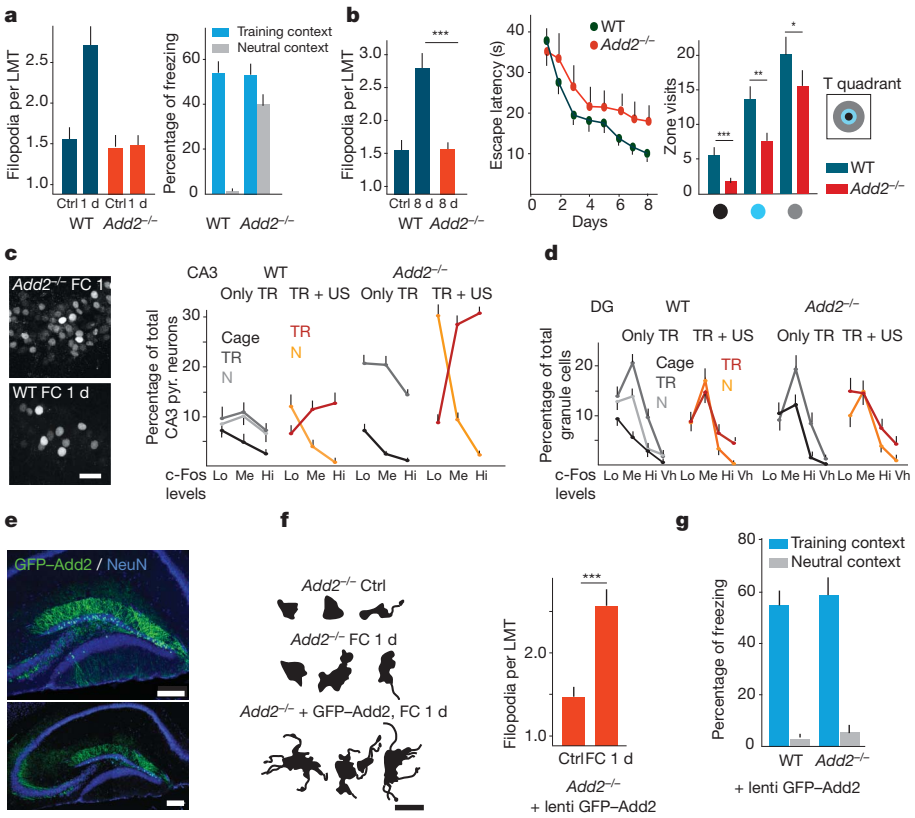


Figure 5 | Critical role of mossy fibre Add2 for feedforward inhibition growth at LMTs and hippocampal memory precision. **a**, Absence of feedforward inhibition growth upon contextual fear conditioning, and generalization in *Add2*$^{-/-}$ mice. Conditions as in Fig. 4c. **b**, Absence of feedforward inhibition growth upon Morris water maze learning, and imprecise spatial memory in *Add2*$^{-/-}$ mice. Conditions as in Fig. 1c. **c**, **d**, c-Fos immunoreactivity in CA3 pyramidal neurons (**c**) and in dentate gyrus (DG) granule cells (**d**) of wild-type (WT) and *Add2*$^{-/-}$ mice upon exposure to training context (TR) or neutral context (N), with or without aversive conditioning. Hi, high; Lo, low; Me, medium; Vh, very high. Conditions as in Fig. 4a. **e**–**g**, Rescue of feedforward inhibition growth and contextualization upon re-expression of Add2 in granule cells of adult *Add2*$^{-/-}$ mice. **e**, Examples of transduced hippocampus (dorsal third of hippocampus). **f**, Rescue of feedforward inhibition growth; lucidas: transgene expression visualized by the GFP–Add2 construct in the absence of mGFP reporter. **g**, Behavioural rescue of contextualization. Conditions as in **a**. Scale bars: 5 μm (**f**), 20 μm (**c**) and 200 μm (**e**). Error bars show mean ± s.e.m.

inhibition connectivity upon learning by mossy fibre LMTs in CA3 is critically important for the behavioural precision of learning-related hippocampal spatial memories. We further show that, upon learning, the increased feedforward inhibition connectivity is brought about through structural plasticity at a substantial fraction of LMTs in CA3, leading to about a doubling in the numbers of excitatory synapses onto parvalbumin-positive inhibitory interneurons (see also Supplementary Material).

Our results introduce a distinction between spatial learning, which is present in *Add2*$^{-/-}$ mice, and the behavioural precision of the learning, which is compromised in these mutant mice. The distinction is consistent with the notion that the hippocampus is critically important for the precision of contextual memorie[25]. Within the hippocampal circuit, the dentate gyrus establishes fine-grained representations of experience, which it transmits to CA3 (ref. 13). Upon learning-induced potentiation, this high-resolution information may augment the detection of similarities among unrelated events through the associational network in CA3. Accordingly, filtering of the mossy fibre output through feedforward inhibition connectivity upon learning[26–28] may support memory precision by restricting the extraction of relational representations in CA3 (ref. 29). The increase in feedforward inhibition connectivity through structural plasticity discovered in this study may thus have important roles in ensuring the precision of behaviourally relevant memories upon learning, under normal and pathological conditions.

METHODS SUMMARY

Rab3a$^{-/-}$ and *Add2*$^{-/-}$ mice[9,10] were from Jackson Laboratories; the reporter line *Thy1-mGFP(Lsi1)* was as described before[5]. The membrane-targeted green fluorescent protein (mGFP) lentivirus to trace mossy fibre projections was as described previously[15]; the GFP–Add2 construct was cloned into a lentivirus vector, and dentate gyrus infections were as described previously[15].

For anatomical analysis, mice were perfused with ice-chilled 4% paraformaldehyde in 0.1 M PBS, and brains were post-fixed. Hippocampi were mounted in 3% agarose blocks, and 100-μm transversal sections of hippocampi were cut using a McIlwain tissue chopper. Sections analysed were within 15% and 30% along the anterior–posterior axis. All LMTs that could be resolved in three dimensions within any given optical field (×100) were analysed for filopodial contents. Filopodia were defined as processes emanating from LMTs of at least 2 μm length; varicosities were defined as end-swellings of at least 1 μm in diameter.

The immuno-electron microscopy analysis was performed according to a published procedure[30].

For c-Fos analysis, mice were perfused for 90 min after the last memory recall. Quantitative analysis of Bassoon puncta and c-Fos-positive nuclei was performed using a computerized image analysis system (Imaris 7, Bitplane). Nuclei were detected automatically as spheres of 8 μm, and the software yielded distributions of c-Fos-positive nuclei. Intensity thresholds for CA3 were defined as follows: low (>280, <450), medium (>450, <700), high (>700; the highest values were about 1,400).

Statistical analyses were performed using Student's *t*-tests and one-way ANOVA; post hoc comparisons were at the $P < 0.05$ level of significance. Results are presented as mean ± s.e.m.

All behavioural experiments were carried out with male mice that were 55–65 days old at the onset of the experiment, and were according to standard procedures. All subsequent morphological and immunohistochemical analyses of behaviourally treated mice were carried out blind to behavioural conditions.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 16 September 2010; accepted 17 February 2011.
Published online 1 May 2011.

1. Holtmaat, A. & Svoboda, K. Experience-dependent structural plasticity in the mammalian brain. *Nature Rev. Neurosci.* **10,** 647–658 (2009).
2. Hübener, M. & Bonhoeffer, T. Searching for engrams. *Neuron* **67,** 363–371 (2010).
3. Lamprecht, R. & LeDoux, J. Structural plasticity and memory. *Nature Rev. Neurosci.* **5,** 45–54 (2004).
4. Wilbrecht, L., Holtmaat, A., Wright, N., Fox, K. & Svoboda, K. Structural plasticity underlies experience-dependent functional plasticity of cortical circuits. *J. Neurosci.* **30,** 4927–4932 (2010).
5. De Paola, V., Arber, S. & Caroni, P. AMPA receptors regulate dynamic equilibrium of presynaptic terminals in mature hippocampal networks. *Nature Neurosci.* **6,** 491–500 (2003).
6. Hofer, S. B., Mrsic-Flogel, T. D., Bonhoeffer, T. & Hübener, M. Experience leaves a lasting structural trace in cortical circuits. *Nature* **457,** 313–317 (2009).
7. Xu, T. *et al.* Rapid formation and selective stabilization of enduring motor memories. *Nature* **462,** 915–919 (2009).
8. Yang, G., Pan, F. & Gan, W. B. Stably maintained dendritic spines are associated with lifelong memories. *Nature* **462,** 920–924 (2009).
9. Castillo, P. E. *et al.* Rab3A is essential for mossy fibre long-term potentiation in the hippocampus. *Nature* **388,** 590–593 (1997).
10. Rabenstein, R. L. *et al.* Impaired synaptic plasticity and learning in mice lacking β-adducin, an actin-regulating protein. *J. Neurosci.* **25,** 2138–2145 (2005).
11. Wang, S.-H. & Morris, R. G. Hippocampal-neocortical interactions in memory formation, consolidation, and reconsolidation. *Annu. Rev. Psychol.* **61,** 49–79 (2010).
12. Nakashiba, T., Young, J. Z., McHugh, T. J., Buhl, D. L. & Tonegawa, S. Transgenic inhibition of synaptic transmission reveals role of CA3 output in hippocampal learning. *Science* **319,** 1260–1264 (2008).
13. Leutgeb, J. K., Leutgeb, S., Moser, M. B. & Moser, E. I. Pattern separation in the dentate gyrus and CA3 of the hippocampus. *Science* **315,** 961–966 (2007).
14. Gogolla, N., Galimberti, I., Deguchi, Y. & Caroni, P. Wnt signaling mediates experience-related regulation of synapse numbers and mossy fiber connectivities in the hippocampus. *Neuron* **62,** 510–525 (2009).
15. Galimberti, I., Bednarek, E., Donato, F. & Caroni, P. EphA4 signaling in juveniles establishes topographic specificity of structural plasticity in the hippocampus. *Neuron* **65,** 627–642 (2010).
16. Acsady, L., Kamondi, A., Sik, A., Freund, T. & Buszaki, G. GABAergic cells are the major postsynaptic target of mossy fibers in the rat hippocampus. *J. Neurosci.* **18,** 3386–3403 (1998).
17. Lawrence, J. J. & McBain, C. J. Interneuron diversity series: containing the detonation—feedforward inhibition in the CA3 hippocampus. *Trends Neurosci.* **26,** 631–640 (2003).
18. Mori, M., Abegg, M. H., Gaehwiler, B. H. & Gerber, U. A frequency-dependent switch from inhibition to excitation in a hippocampal unitary circuit. *Nature* **431,** 453–456 (2004).
19. D'Angelo, E. & De Zeeuw, C. I. Timing and plasticity in the cerebellum: focus on the granular layer. *Trends Neurosci.* **32,** 30–40 (2009).
20. Sacchetti, B., Scelfo, B., Tempia, F. & Strata, P. Long-term synaptic changes induced in the cerebellar cortex by fear conditioning. *Neuron* **42,** 973–982 (2004).
21. Wiltgen, B. J. & Silva, A. J. Memory for context becomes less specific with time. *Learn. Mem.* **14,** 313–317 (2007).
22. Biedenkapp, J. C. & Rudy, J. W. Context pre-exposure prevents forgetting of a contextual fear memory: implication for regional changes in brain activation patterns associated with remote and recent memory tests. *Learn. Mem.* **14,** 200–203 (2007).
23. Kubik, S., Miyashita, T. & Guzowski, J. F. Using immediate-early genes to map hippocampal subregional functions. *Learn. Mem.* **14,** 758–770 (2007).
24. Bednarek, E. & Caroni, P. β-Adducin is required for stable assembly of new synapses and improved memory upon environmental enrichment. *Neuron.* (in the press).
25. Wiltgen, B. J. *et al.* The hippocampus plays a selective role in the retrieval of detailed contextual memories. *Curr. Biol.* **20,** 1336–1344 (2010).
26. Lamsa, K., Heeroma, J. H. & Kullmann, D. M. Hebbian LTP in feed-forward inhibitory interneurons and the temporal fidelity of input discrimination. *Nature Neurosci.* **8,** 916–924 (2005).
27. Wulff, P. *et al.* Synaptic inhibition of Purkinje cells mediates consolidation of vestibulo-cerebellar motor learning. *Nature Neurosci.* **12,** 1042–1049 (2009).
28. Pouille, F., Marin-Burgin, A., Adesnik, H., Atallah, B. V. & Scanziani, M. Input normalization by global feedforward inhibition expands cortical dynamic range. *Nature Neurosci.* **12,** 1577–1585 (2009).
29. McNaughton, B. L. & Morris, R. G. M. Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends Neurosci.* **10,** 408–415 (1987).
30. Knott, G. W., Holtmaat, A., Trachtenberg, J. T., Svoboda, K. & Welker, E. A protocol for preparing GFP-labeled neurons previously imaged *in vivo* and in slice preparations for light and electron microscopic analysis. *Nature Protocols* **4,** 1145–1156 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Arber and B. Roska for valuable comments on the manuscript. We are grateful to J. Pielage for sharing with us his findings on the function of Add2 in synapse stability, and to G. Courtine for advice on the c-Fos labelling protocol. The Friedrich Miescher Institut is part of the Novartis Research Foundation.

Author Contributions S.R. devised, carried out and analysed all experiments except for those of Fig. 2a–c, part of Fig. 2d, Fig. 5a, e–g and Supplementary Fig. 4; C.V. carried out the experiments of Fig. 2a–c, part of Fig. 2d and Supplementary Fig. 4; E.B. devised and carried out the behavioural and rescue experiments on *Add2*$^{-/-}$ mice; C.G. carried out the immuno-electron microscopy experiments; B.S. provided advice in planning and interpreting the fear conditioning experiments; P.S. provided advice on the cerebellar experiments; P.C. helped devise the experiments and wrote the manuscript. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to P.C (caroni@fmi.ch).

METHODS

Reagents and immunocytochemistry. Antibodies were from the following sources, and were used as follows: parvalbumin, Swant, 1:5,000; VGluT1, SySy, 1:1,000; GAD65/67, Millipore, 1:1,000; c-Fos, Santa Cruz, 1:10,000; NeuN, Chemicon, 1:200; Bassoon, Millipore, 1:200; Alexa-labelled secondary antibodies, Molecular Probes, 1:500.

For immunocytochemistry, tissues were permeabilized with 0.2% Triton X-100 in PBS with 10% bovine serum albumin (BSA). Antibody incubations were overnight at 4 °C.

Fluorescence was imaged on either an upright spinning disk microscope consisting of a Yokogawa CSU22 confocal scanning head mounted on a Zeiss Axioimager M1 using a ×100 alphaPlan-Apochromat 1.45 (Zeiss) oil-immersion objective, or on an LSM510 confocal microscope (Zeiss) using a ×63 (1.4) oil-immersion objective.

At least four sections were analysed per mouse, and the data are based on 300–500 μm regions along the anterior–posterior axis.

c-Fos analysis. For c-Fos analysis, all samples belonging to the same experimental set were processed in parallel. Occasional sections in which NeuN signals were lower than average, or where c-Fos signal intensities varied within different regions of the section were discarded as technically poor. All images were acquired with the same settings, which were defined in order to avoid saturation of the highest c-Fos signals in CA3 and dentate gyrus, and to still detect background levels outside cell clusters. Cells were binned according to labelling intensities using an automatic procedure, and the same threshold settings were used for all experiments. For dentate gyrus granule cells, the thresholds were as follows: low (>280, <450), medium (>450, <700), high (>700, <1,000), very high (>1,000; the highest values were about 2,200). c-Fos immunoreactive neurons were counted using a minimum of four sections per animal, and normalized to the total number of NeuN-positive nuclei within the neuronal layers in CA3 or dentate gyrus. In a first series of experiments, batches of naive and fear conditioning control mice (training context without unconditioned aversive stimulus) were tested for inter-animal variability, which was found to be very low.

Behavioural experiments. The behavioural experiments were in accordance with institutional guidelines, and were approved by the Veterinary Department of the Canton of Basel-Stadt. Mice were kept in temperature-controlled rooms on a constant 12 h light/dark cycle, and experiments were conducted at the approximate same time during the light cycle. Before the behavioural experiments, mice were kept in a holding room in single cages for 3–4 days. At the onset of each behavioural experiment mice were 50–60 days old.

For the Morris water maze test, the 140 cm pool was surrounded by black curtains, and by four different objects. A circular escape platform (10 cm diameter) was submerged 0.5 cm below the water surface, and was kept in a fixed position. Mice were trained to find the platform for 4 trials a day, during up to 8 days. During training, mice were released from pseudo-randomly assigned start locations; they were allowed to swim for up to 60 s, when they were manually guided to the platform in the case of failures. Inter-trial intervals were 5 min. Single probe trials to test reference memory were conducted 1 day after the last training session. Mice were released at a random start position, and were allowed to swim during 60 s in the absence of the platform.

The training context (TR) was rectangular, and was cleaned with 1% acetic acid before and after each trial; the neutral context (N) had a cylindrical shape and was cleaned with 70% ethanol. Freezing was defined as the absence of somatic motility, except for respiratory movements. Exploratory activity was measured as body distance travelled over time. Once placed in the conditioning chamber, the mice were allowed to freely explore for 2.5 min, and they received 5 presentation of conditioned stimulus and unconditioned stimulus (1 s foot shock, 0.8 mA; where indicated, 10 kHz tone for 10 s, 70 dB sound pressure level, inter-trial interval 30 s). The last 1 s of each tone was paired with the unconditioned stimulus. Contextual fear conditioning involved the same protocol, but without the tone component. To test for contextual fear memory, mice were returned to training (or neutral) context during a test period of 2.5 min. To test for cued fear conditioning, mice explored for 2 min, followed by 5 tone presentations. The test was performed either in the conditioning context (context- and tone-dependent freezing), or in a novel context (tone-dependent freezing).

To test for context discrimination after fear conditioning, a within-subjects design was used. On the test day, freezing was assessed in training context during 2.5 min, and 5 h later in neutral context. Where indicated, mice were tested for generalization in neutral context, followed 5 h and 24 h later by two brief recall sessions (in training or neutral context). Subsequently, discrimination was tested in a second novel context (novel room shape; 0.25% benzaldehyde/ethanol).

Data from training sessions and probe trials were collected and analysed using Viewer2 Software (Biobserve). Cued and contextual fear conditioning were carried out in the Mouse Test Cage (Coulbourn Instruments). Freezing behaviour was scored using Ethovision software (Noldus). Mice were excluded from the data set if they failed at the behavioural analysis; this was the case when mice failed to extinguish fear responses to training context (two mice), exhibited weak freezing to training context in the recall experiments at day 15 (three mice), exhibited signs of behavioural extinction upon recall of training context at day 15 (seven mice), or failed to learn the Morris water maze (one mouse).

Transmission electron microscopy. This procedure is described in detail elsewhere[30]. Briefly, mice were transcardially perfused with 2% paraformaldehyde and 0.2% glutaraldehyde in PBS 0.1 M pH 7.4. Right and left hippocampi were dissected, and 60 μm vibratome (Leica) sections were obtained, rinsed, cryoprotected and freeze-thawed in liquid nitrogen. Sections were incubated in first antibody (GFP, chemicon 1:1,000) overnight, followed by biotinylated secondary antibody (Invitrogen 1:500). After incubation in the avidin-biotin peroxidase complex (ABC elite, Vector Laboratories), labelling was performed with DAB and hydrogen peroxide. After the revelation of the labelling, sections were stained in osmium tetroxide and dehydrated. After impregnation with Durcupan resin (FLUKA) sections were flat-embedded between two silicon-coated glass slides and cured in a 60 °C oven for 48 h.

Transmission light microscopy was performed in stratum lucidum to search for large mossy fibre terminals with more than three filopodia. Appropriate blocks were then trimmed, and 60 nm serial sections were cut and collected on formvar coated slot-grids. Images of labelled terminal were acquired with a side-mounted digital camera (Veleta, Olympus) on a Philips CM10 transmission electron microscopy at 80 kV, and a pixel size of 2.63 nm. To reconstruct the structure in three dimensions, images were aligned (Autoaligner, Bitplane), and contours were drawn manually using Imaris 7.1.2 (Bitplane). Surface rendering was achieved using Geometry converter (J. Wolf) and Blender.

LETTER

doi:10.1038/nature10004

Long-term evolution and transmission dynamics of swine influenza A virus

Dhanasekaran Vijaykrishna[1,2,3], Gavin J. D. Smith[1,2,3], Oliver G. Pybus[4], Huachen Zhu[1,2], Samir Bhatt[4], Leo L. M. Poon[1], Steven Riley[5], Justin Bahl[1,2,3], Siu K. Ma[1], Chung L. Cheung[1], Ranawaka A. P. M. Perera[1], Honglin Chen[1,2], Kennedy F. Shortridge[1,2], Richard J. Webby[6], Robert G. Webster[1,6], Yi Guan[1,2] & J. S. Malik Peiris[1,7]

Swine influenza A viruses (SwIV) cause significant economic losses in animal husbandry as well as instances of human disease[1] and occasionally give rise to human pandemics[2], including that caused by the H1N1/2009 virus[3,4]. The lack of systematic and longitudinal influenza surveillance in pigs has hampered attempts to reconstruct the origins of this pandemic[4]. Most existing swine data were derived from opportunistic samples collected from diseased pigs in disparate geographical regions, not from prospective studies in defined locations, hence the evolutionary and transmission dynamics of SwIV are poorly understood. Here we quantify the epidemiological, genetic and antigenic dynamics of SwIV in Hong Kong using a data set of more than 650 SwIV isolates and more than 800 swine sera from 12 years of systematic surveillance in this region, supplemented with data stretching back 34 years. Intercontinental virus movement has led to reassortment and lineage replacement, creating an antigenically and genetically diverse virus population whose dynamics are quantitatively different from those previously observed for human influenza viruses. Our findings indicate that increased antigenic drift is associated with reassortment events and offer insights into the emergence of influenza viruses with epidemic potential in swine and humans.

All major SwIV lineages of North American or European origin—classical swine (CS), European or Eurasian avian-like swine (EA) and triple-reassortant swine (TRIG) (see Supplementary Information for an overview)—co-circulate in southern China[4,5], and both human (H3N2) and avian (H5N1 and H9N2) viruses have been isolated from swine in the region[6–9]. To address the critical lack of structured swine influenza data, we undertook virological and serological analysis of Hong Kong SwIV surveillance samples. Most (80–95%) of the swine slaughtered in Hong Kong originate from provinces in mainland China (Supplementary Fig. 1 and Supplementary Table 1), the region with the world's largest swine population[10–12].

We isolated and subtyped 573 H1N1 and H1N2, 97 H3 and 2 H9N2 viruses from fortnightly sampling of swine slaughtered between May 1998 and January 2010 (Fig. 1a, b). We found no H5N1 viruses. From August 1998 to December 2002, the isolates were mostly CS H1N1 viruses. EA H1N1 viruses were detected only from 2001 onwards and TRIG H1N2 from 2002 onwards. During 2002–05, viruses classified as CS, EA, TRIG and H3N2 co-circulated and fluctuated in relative prevalence (Fig. 1b). After 2005, EA H1N1 viruses became dominant and H3N2 viruses disappeared, although CS H1N2 and TRIG H1N2 viruses continued to be isolated sporadically (Fig. 1b). All three SwIV H1 lineages (CS, EA and TRIG) have co-circulated with H1N1/2009 after the introduction of the latter virus into pigs[5].

Comprehensive phylogenetic analyses of the genes encoding surface antigens haemagglutinin (HA) and neuraminidase (NA) in all H1N1 and H1N2 isolates (including 93 H1N1 viruses isolated during 1976–79 and 1993–94) confirmed that most isolates belonged to the CS or EA lineages (Fig. 2 and Supplementary Fig. 2a–c). All pre-1998 viruses were CS except two 'pure avian' viruses from 1993 (ref. 13); much greater HA diversity was observed after 1998 (Fig. 2a). Notably, our CS isolates do not form a single monophyletic group; rather, they are interspersed with North American CS viruses, indicating multiple introductions of CS into the study area. In contrast, EA viruses are monophyletic, indicating a single introduction. All Hong Kong TRIG viruses form a single group

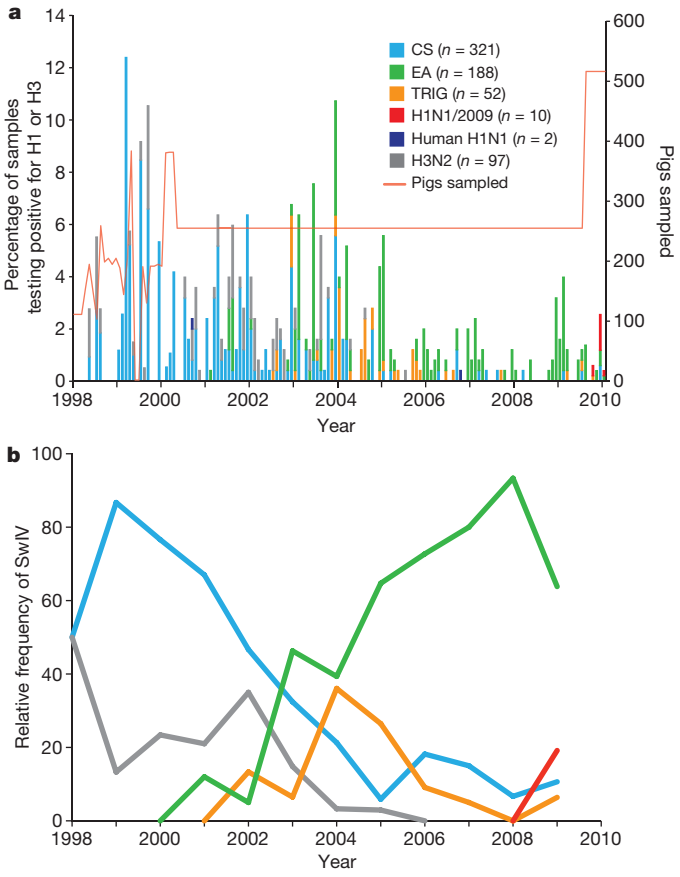


Figure 1 | Prevalence and relative frequency of swine influenza H1 and H3 subtypes. a, **b**, Percentage prevalence (**a**) and year-averaged relative frequency (**b**) of the major HA variants of SwIV. Colour codes and numbers of isolates (*n*) of H1 and H3 subtype viruses detected from swine in Hong Kong between 1998 and 2010 are shown. The viruses detected include CS, EA, TRIG, H1N1/2009 and human seasonal H1N1 and H3N2 viruses. The orange line indicates the number of pigs sampled during the surveillance period.

[1]State Key Laboratory of Emerging Infectious Diseases & Department of Microbiology, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 21 Sassoon Road, Pokfulam, Hong Kong, China. [2]International Institute of Infection and Immunity, Shantou University Medical College, Shantou, Guangdong, China. [3]Laboratory of Virus Evolution, Program in Emerging Infectious Diseases, Duke-NUS Graduate Medical School, 8 College Rd, 169857, Singapore. [4]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK. [5]Department of Community Medicine and School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong, China. [6]Virology Division, Department of Infectious Diseases, St Jude Children's Research Hospital, Memphis, Tennessee 38015, USA. [7]HKU-Pasteur Research Centre, The University of Hong Kong, Pokfulam, Hong Kong Special Administrative Region, China.

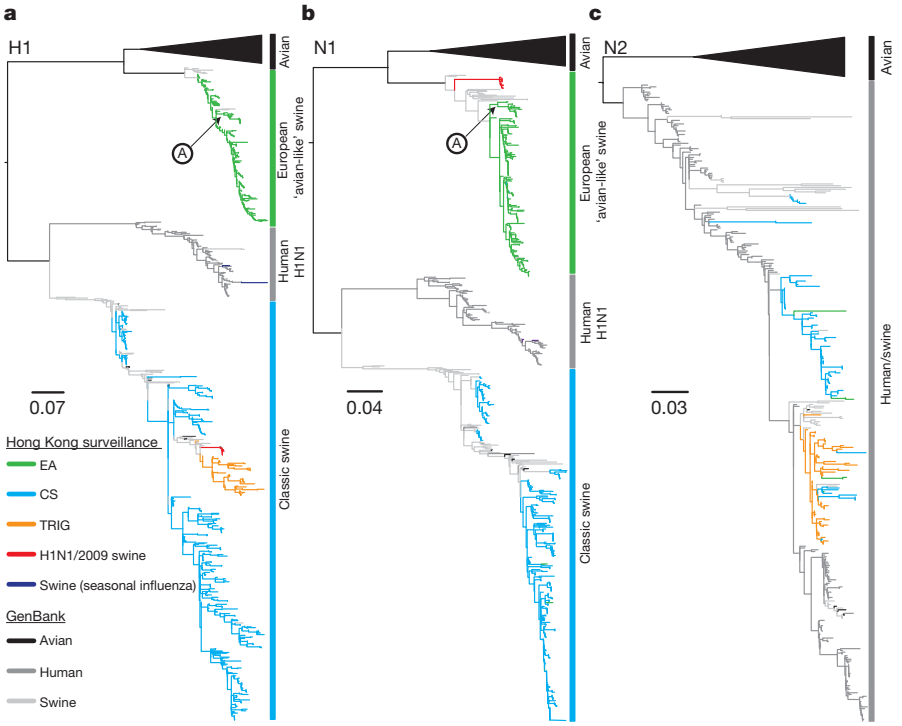


Figure 2 | Genetic relationships of swine influenza A viruses for genes encoding surface proteins. **a**–**c**, Haemagglutinin H1 (**a**), neuraminidase N1 (**b**), neuraminidase N2 (**c**). Representative avian, swine and human influenza A viruses obtained from GenBank are represented by black, dark grey and light grey, respectively. Colour codes of H1N1 and H1N2 subtype viruses detected from swine during 1977–2010 in Hong Kong are shown in the key. Arrow A indicates the antigenically divergent EA-reassortant viruses discussed in the text (for example, Sw/HK/72/2007, Fig. 4). Scale bars represent substitutions per site. Fully detailed phylogenies including sequence names are provided in Supplementary Fig. 2a–c.

except for isolate Sw/HK/78/2003, indicating that this virus was introduced from North America separately. The Hong Kong TRIG viruses diverge from the North American TRIG and H1N1/2009 HA lineages soon after the emergence of TRIG H1N2 viruses in North America and thus constitute a third distinct TRIG HA gene lineage (Fig. 2a)[14].

Molecular clock phylogenies of 221 whole genomes (33.2% of isolates) revealed that CS viruses isolated from 1976 to 1994 contained only CS genome segments: no reassortment was detected during this period (Fig. 3a and Supplementary Fig. 3). However, several reassortant SwIV were detected between 1998 and 2010. In addition to the CS, EA, TRIG and H1N1/2009 viruses, we detected 14 'genotypes' generated by reassortment between circulating swine and human/avian lineages (Supplementary Figs 3 and 4). Most of the newly identified reassortants were observed only transiently and usually contained genome segments from viruses that were dominant at that time (Supplementary Fig. 4). Excepting the CS H1N2 virus, which acquired the human H3N2 neuraminidase gene repeatedly, we detected no preferential direction of horizontal gene transfer among SwIV strains.

Three of the 14 reassortant genotypes were isolated repeatedly (Supplementary Fig. 4); specifically, (1) CS H1N2 viruses, (2) novel H1N2 reassortants and (3) Sw/HK/72/2007-like (EA-reassortant) strains, which have acquired an NS gene from TRIG viruses and which belong to a divergent EA lineage (Fig. 2). Since their initial detection in 2007, EA-reassortant viruses have become the dominant EA lineage, constituting 12.5% of all EA viruses in 2007, 15.4% in 2008 and 41.4% in 2009. Because most reassortant 'genotypes' were isolated only once, we hypothesize that few are adapted for continuous circulation (although we cannot exclude stochastic demographic effects or sampling bias as alternative explanations). SwIV diversity in our population is probably increased by pig movements: breeding pigs constitute the bulk of live pigs imported into China and data indicate that imports have increased since 1990 (refs 11,12).

For all genome segments, molecular clock phylogenies exhibited long branches leading to several reassortant lineages (Fig. 3). This was also observed for the H1N1/2009 virus[4] and indicates a long period of unsampled diversity. Upon first detection, these reassortant lineages tend to be more closely related to viruses circulating in our population five to eleven years previously, rather than to co-circulating strains (that is, they do not arise from the contemporaneous part of the phylogenetic 'backbone').

We found extensive antigenic crossreaction among CS, TRIG and H1N1/2009 viruses (Supplementary Table 2 and Supplementary Fig. 5). Ferret antisera to these viruses also crossreacted with early EA viruses (2001–03) but reacted poorly with more recent EA-reassortant strains (Fig. 4). Interestingly, the six novel EA-reassortant viruses tested (sampled between 2007 and 2009) crossreacted weakly with all ferret antisera used, including the antiserum to Sw/HK/NS29/2009. This group thus represents an antigenically distinct SwIV lineage. Excepting the earliest EA reassortant (Sw/HK/72/2007), all remaining EA reassortants reacted well against antisera raised to the EA-reassortant virus Sw/HK/1559/2009 (Fig. 4), indicating progressive antigenic change of EA-reassortant viruses during our study.

The earliest of the above-mentioned EA reassortants (Sw/HK/72/2007) had acquired two amino acid changes in HA antigenic sites and later EA reassortants (for example, Sw/HK/1559/2008 and Sw/HK/1532/2009) had a further five changes at antigenic sites (Supplementary Figs 3a and 6). These findings support the hypothesis that EA-reassortant viruses have antigenically drifted away from crossreacting antibodies arising from CS, TRIG and early-EA virus infection. The observation that antigenic change occurred in the reassortant EA virus lineage rather than in the parental lineage indicates that reassortment may facilitate the generation of SwIV antigenic diversity.

Although SwIV isolation rates declined after EA viruses became predominant (Fig. 1), serological data indicate that overall SwIV seroprevalence has not declined (Supplementary Tables 3 and 4). To test whether EA viruses have a competitive advantage over CS strains, we intranasally infected five-week-old, previously influenza-naive pigs with SwIV representative of the lineages isolated here (Methods and

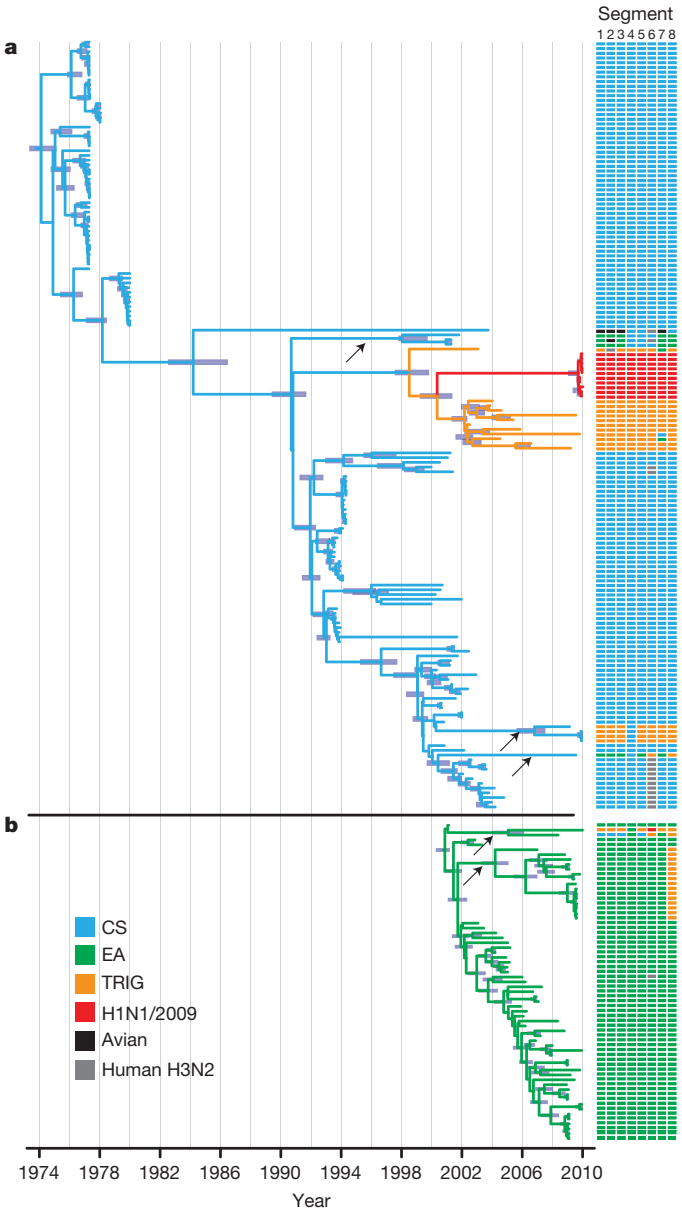


Figure 3 | Phylogenies and divergence times of the haemagglutinin genes of classical swine and European avian-like SwIV. **a**, CS; **b**, EA. Coloured boxes adjacent to tips show the lineage classification of each gene segment of SwIV isolated in this study. Arrows indicate the long branches that lead to newly detected reassortant SwIV. Purple node bars represent 95% credible intervals of lineage divergence times. A fully detailed HA phylogeny including sequence names is shown in Supplementary Fig. 3a.

Test antigens		Ferret antisera: 4167/1999 CS H1N1	1304/2003 CS H1N2	1110/2006 TRIG H1N2	Cal/4/2009 H1N1/09	NS29/2009 EA H1N1	1559/2008 EA* H1N1
Sw/HK/4167/1999	CS H1N1	1:20,480	1:320	1:10,240	1:2,560	1:2,560	<1:10
Sw/HK/1304/2003	CS H1N2	1:1,280	1:2,560	1:640	1:80	1:40	<1:10
Sw/HK/1110/2006	TRIG H1N2	1:40,960	1:1,280	1:10,240	1:640	1:5,120	<1:10
Cal/04/2009	H1N1/2009	1:640	1:640	1:2,560	1:1,280	160	<1:10
Sw/HK/8512/2001	EA H1N1	1:10,240	1:1,280	1:5,120	1:2,560	1:10,240	1:320
Sw/HK/1669/2002	EA H1N1	1:5,120	1:1,280	1:2,560	1:1,280	1:10,240	1:160
Sw/HK/NS129/2003	EA H1N1	1:5,120	1:1,280	1:2,560	1:1,280	1:10,240	1:160
Sw/HK/1716/2006	EA H1N1	1:2,560	1:640	1:1,280	1:640	1:2,560	<1:10
Sw/HK/NS952/2008	EA H1N1	1:2,560	1:640	1:640	1:320	1:2,560	<1:10
Sw/HK/NS29/2009	EA H1N1	1:640	1:160	1:1,280	1:80	1:10,240	<1:10
Sw/HK/72/2007	EA* H1N1	<1:10	<1:10	<1:10	<1:10	1:40	<1:10
Sw/HK/1559/2008	EA* H1N1	<1:10	<1:10	<1:10	<1:10	1:40	1:5,120
Sw/HK/247/2009	EA* H1N1	<1:10	<1:10	<1:10	<1:10	1:40	1:2,560
Sw/HK/NS613/2009	EA* H1N1	<1:10	<1:10	<1:10	<1:10	1:10	1:2,560
Sw/HK/2481/2009	EA* H1N1	<1:10	<1:10	<1:10	<1:10	1:20	1:2,560
Sw/HK/NS186/2009	EA* H1N1	<1:10	<1:10	<1:10	<1:10	1:40	1:2,560

Figure 4 | Antigenic characterization of SwIV measured by haemagglutinin inhibition assays. Titres are shaded according to their respective major SwIV HA lineages (see Figs 1–3); low (1:20, 1:40) and non-reactive titres (<1:10) are shaded in lighter colours. Underlined values represent homologous antibody titres. EA-reassortant viruses (indicated by asterisks) showed poor crossreactivity against antisera raised towards CS, TRIG, H1N1/2009 and late (2006–2009) 'pure' EA viruses. Excepting the earliest EA-reassortant virus (Sw/HK/72/2007), all remaining EA-reassortants reacted well against antisera raised towards the EA-reassortant virus Sw/HK/1559/2009, indicating progressive antigenic change of this novel reassortant.

Supplementary Fig. 7). EA viruses showed the highest and most prolonged virus shedding, closely followed by TRIG viruses; CS viruses showed lower peak viral titres. Thus, the replicative advantage of EA viruses, together with the low prevalence of crossreactive antibodies to EA in swine (15% in 2000, 26% in 2004; Supplementary Table 4), may help to explain the replacement of other SwIV lineages with EA viruses.

We tracked the evolution in our EA viruses of amino acids previously associated with adaptation of avian influenza to other species[15–17]. Purported avian residues were maintained at most of these sites (Supplementary Table 5) despite the circulation of these viruses in swine for more than 30 years[18,19]. However, the PDZ- (post-synaptic density protein, *Drosophila* disc large tumor suppressor and zonula occludens-1 protein) ligand at the 3′ end of EA virus non-structural (NS) 1 genes showed significant host-specific evolution: early European EA viruses had the avian ESEV motif, with a change to GSEV/GPEV motifs observed in several hosts. By 1999, most viruses sampled had the GPKV motif previously described from pigs[16]. CS and TRIG viruses that contributed the NS gene to H1N1/2009 have a truncated NS gene, as do the antigenically variant Sw/HK/72/2007-like viruses. The role of the truncated NS gene in inter-species transmission clearly merits further study. Furthermore, a modest but significant ($P < 0.01$) change in selection pressure was observed between European EA viruses isolated shortly after cross-species transmission (non-synonymous to synonymous (d_N/d_S) substitution rate ratio of 0.24; 95% confidence interval = 0.22–0.27) and those isolated later ($d_N/d_S = 0.17$; 95% confidence interval = 0.14–0.20), consistent with the hypothesis that host-specific selection increased viral adaptation after the introduction of EA viruses into swine (Supplementary Table 6).

Our unique longitudinal study reveals a genetically and antigenically dynamic SwIV population within a single region and provides a baseline for future studies of the virus elsewhere. The epidemiology and evolution of SwIV seem to be strongly shaped by gene flow among continents and species, facilitating the reassortment of diverse lineages and occasionally resulting in antigenic change. Although we confirm that the H1N1/2009 virus was not generated within our study's catchment, the processes of lineage emergence, importation, reassortment and replacement described here are probably representative of the H1N1/2009 source population. We show that reassortments between EA and TRIG viruses do occur, generating reassortants that establish themselves as stable lineages in swine. SwIV reassortants containing H1N1/2009-like genome segments have also been transiently detected[5,20].

Despite clear evidence of inter-continental SwIV movement, gene flow is not so frequent that the global SwIV population acts as a single gene pool (as observed for human influenza A[21,22]); instead a higher diversity of mammalian-adapted viruses in global swine populations is supported. Crucially, the co-circulation of multiple SwIV lineages facilitates the production of new genomic combinations. The evolutionary consequences of increased SwIV movement are hard to predict but require consideration given an increasingly globalized future.

Our study reveals a frequent generation of new reassortants but the survival and persistence of only a few, a process we term 'recombinant chatter'[23]. Our data also indicate that reassortment and antigenic change are linked. This phenomenon was described in North America CS viruses after the events that generated the TRIG viruses[24]; it has also been observed in human influenza[21]. After reassortment, evolution in HA antigenic domains may arise for several reasons: (1) because of herd-immune selection pressure; (2) because those residues are under weak selective constraint; or (3) to compensate for fitness costs of mutations accruing elsewhere in the genome. The role of reassortment in driving genome-wide evolution requires detailed investigation.

We found that the quantitative dynamics of SwIV genomic diversity and lineage turnover (Supplementary Fig. 8) are slower, less periodic and less predictable than the repeated annual replacements typically seen for human influenza A. The reasons for SwIV lineage change are unclear: previously, selection arising from herd immunity was considered less important for pigs than for humans because the short lifespan of farmed swine (~150 days) lowers the chance of re-infection, reducing the cross-protection that probably drives antigenic drift. Furthermore, maternally acquired swine immunity does not seem to interrupt SwIV infection or transmission, despite masking clinical illness[25]. Our surveillance data, animal infection experiments and serological data show that one reason for lineage change may be a competitive advantage of EA over CS and TRIG viruses.

The hypothesis that pigs are important in pandemic emergence, as facilitators of reassortment among influenza viruses[26], has regained favour after the emergence of H1N1/2009. This virus represented a subtype already endemic in humans, implying that other H1 and H3 viruses prevalent in swine are credible pandemic candidates, especially when corresponding immunity in humans is absent. Indeed, we found that there are other swine viruses (for example, Sw/HK/NS29/09) to which humans lack herd immunity (Supplementary Table 7). Hence, future assessment of zoonotic potential must combine the evaluation of crossreactive immunity in humans, the assessment of transmissibility in animal models and ongoing surveillance of SwIV genetic diversity. The H1N1/2009 virus has already infected swine and reassorted with other SwIV, indicating that circulating SwIV will continue to acquire novel non-SwIV genes[5] (notably, avian viruses such as H9N2 and H5N1 are occasionally detected in swine in Asia[6,8,9,27]). Avian-to-swine and swine-to-human host adaptation of influenza viruses are both poorly understood in comparison to avian-to-human adaptation and are a priority for future research.

METHODS SUMMARY

Systematic swine influenza A virus surveillance was initiated in May 1998 at a central slaughterhouse in Hong Kong. During 1999–2007, about 15–20% of the pigs were farmed locally in Hong Kong and the remainder were imported from several provinces in China; however, since 2008 the proportion of locally produced pigs fell to 5% (see Supporting Information). About 128 nasal and tracheal swabs were collected twice monthly from August 1998 to April 2009; since May 2009, sample numbers were doubled (Fig. 1a). Genes encoding surface proteins (HA and NA) were sequenced for all 573 H1N1 viruses isolated from 1998 to 2010 and for 93 swine H1 viruses from our repository, isolated during the periods 1976–1978 and 1993–1994. Full genome sequencing was carried out for 221 representative viruses. To estimate the genetic diversity and the level of gene reassortment, phylogenetic trees were constructed for each genomic segment independently (Supplementary Fig. 2). On the basis of the phylogenetic relationships of each gene segment, major swine virus lineages circulating in Hong Kong were identified and a more detailed Bayesian phylogenetic analysis for each lineage was conducted, thereby estimating rates of viral evolution and dates of divergence (Supplementary Fig. 3).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 22 September 2010; accepted 17 March 2011.

1. Shinde, V. *et al.* Triple-reassortant swine influenza A (H1) in humans in the United States, 2005–2009. *N. Engl. J. Med.* **360,** 2616–2625 (2009).
2. Smith, G. J. D. *et al.* Dating the emergence of pandemic influenza A viruses. *Proc. Natl Acad. Sci. USA* **106,** 11709–11712 (2009).
3. Garten, R. J. *et al.* Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science* **325,** 197–201 (2009).
4. Smith, G. J. D. *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459,** 1122–1125 (2009).
5. Vijaykrishna, D. *et al.* Reassortment of pandemic H1N1 viruses in swine. *Science* **328,** 1529 (2010).
6. Peiris, J. S. M. *et al.* Cocirculation of avian H9N2 and contemporary "human" H3N2 influenza A viruses in pigs in southeastern China: potential for genetic reassortment? *J. Virol.* **75,** 9679–9686 (2001).
7. Yu, H. *et al.* Genetic evolution of swine influenza A (H3N2) viruses in China from 1970–2006. *J. Virol.* **46,** 1067–1075 (2008).
8. Cong, Y. L. *et al.* Antigenic and genetic characterization of H9N2 swine influenza viruses in China. *J. Gen. Virol.* **88,** 2035–2041 (2007).
9. Li, H. Y. *et al.* Isolation and characterization of H5N1 and H9N2 influenza viruses from pigs in China. *Chinese J. Vet. Prev. Med.* **26,** 1–6 (2004).
10. Clements, A. C. A., Pfeiffer, D. U., Otte, M. J., Morteo, K. & Chen, L. A global livestock production and health atlas (GLiPHA) for interactive presentation, integration and analysis of livestock data. *Prev. Vet. Med.* **56,** 19–32 (2002).
11. Zhang, J. & Beckman, C. People's Republic of China: Agricultural situation: Livestock and Products 2008. (*USDA Foreign Agriculture Service*, 2008).
12. Wang, R. China – pork powerhouse of the world. *Advances Pork Prod.* **17,** 33–46 (2006).
13. Guan, Y. *et al.* Emergence of avian H1N1 influenza viruses in pigs in China. *J. Virol.* **70,** 8041–8046 (1996).
14. Lorusso, A. *et al.* Genetic and antigenic characterization of H1 influenza viruses from United States swine from 2008. *J. Gen. Virol.* **92,** 919–930 (2011).
15. Finkelstein, D. B. *et al.* Persistent host markers in pandemic and H5N1 influenza viruses. *J. Virol.* **81,** 10292–10299 (2007).
16. Obenauer, J. C. *et al.* Large-scale sequence analysis of avian influenza isolates. *Science* **311,** 1576–1580 (2006).
17. Taubenberger, J. K. *et al.* Characterization of the 1918 influenza virus polymerase genes. *Nature* **437,** 889–893 (2005).
18. Pensaert, M., Ottis, K., Vanderputte, J., Kaplan, M. M. & Buchmann, P. A. Evidence for the natural transmission of influenza A virus from wild ducks to swine and its potential for man. *Bull. World Health Organ.* **59,** 75–78 (1981).
19. Dunham, E. *et al.* Different evolutionary trajectories of European avian-like and classical swine H1N1 influenza A viruses. *J. Virol.* **83,** 5485–5494 (2009).
20. Moreno, A. *et al.* Novel H1N2 swine influenza reassortant strain in pigs derived from the pandemic H1N1/2009 virus. *Vet. Microbiol.* **149**, 472–477 (2011).
21. Rambaut, A. *et al.* The genomic and epidemiological dynamics of human influenza A virus. *Nature* **453,** 615–619 (2008).
22. Russell, C. A. *et al.* The global circulation of seasonal influenza A (H3N2) viruses. *Science* **320,** 340–346 (2008).
23. Wolfe, N. D., Dunavan, C. P. & Diamond, J. Origins of major human infectious diseases. *Nature* **447,** 279–283 (2007).
24. Webby, R. J. *et al.* Evolution of swine H3N2 influenza viruses in the United States. *J. Virol.* **74,** 8243–8251 (2000).
25. Loeffen, W. L. A., Heinen, P. P., Bianchi, A. T. J., Hunneman, W. A. & Verheijden, J. H. M. Effect of maternally derived antibodies on the clinical signs and immune response in pigs after primary and secondary infection with an influenza H1N1 virus. *Vet. Immunol. Immunopathol.* **92,** 23–35 (2003).
26. Scholtissek, C., Hinshaw, V. S. & Olsen, C. W. Influenza in pigs and their role as the intermediate host. In *Textbook of Influenza* (eds Nicholson, K. G., Webster R. G. & Hay A. J.) 137–145 (Blackwell Scientific, 1998).
27. Nidom, C. A. *et al.* Influenza A (H5N1) viruses from pigs, Indonesia. *Emerg. Infect. Dis.* **16,** 1515–1523 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This research was supported in part by the National Institute of Allergy and Infectious Diseases (NIAID) contract HHSN26600700005C and the Area of Excellence Scheme of the University Grants Commission (grant AoE/M-12/06) of the Hong Kong SAR Government. We acknowledge the Food and Environmental Hygiene Department of Hong Kong for facilitating the study. We acknowledge support from The Royal Society of London (O.G.P.), UK COSI (S.B.), NIAID (G.J.D.S.), the Agency for Science, Technology and Research and the Ministry of Health, Singapore (D.V., G.J.D.S and J.B.). We thank C. Y. H. Leung for producing some of the ferret antisera used in this study.

Author Contributions J.S.M.P. and Y.G. conceived the study, conducted surveillance, performed analyses and co-wrote the paper. D.V., G.J.D.S. and O.G.P. conceived the study, performed analyses, co-wrote the paper and contributed equally to this work. H.Z., S.B., L.L.M.P., S.R., J.B., R.A.P.M.P. and H.C. performed analyses, S.K.M. conducted surveillance, C.L.C. conducted sequencing, K.F.S. and R.G.W. initiated surveillance in 1976 and provided viruses and R.J.W. provided viruses and reagents. All authors commented on and edited the paper.

Author Information Sequences generated in this study have been deposited with GenBank under the accession numbers CY084470–CY085121, CY085301–CY086876 and CY087041–CY087142. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to Y.G. (yguan@hkucc.hku.hk) or J.S.M.P. (malik@hkucc.hku.hk).

METHODS

Surveillance. Systematic influenza surveillance was conducted from May 1998 until January 2010 in swine at an abattoir in Hong Kong, where tracheal or nasal swabs were collected fortnightly from slaughtered swine. During 1999–2007, about 15–20% of the pigs were farmed locally in Hong Kong and the remainder were imported from several provinces in China; however, since 2008 the proportion of locally produced pigs fell to 5% (see Supporting Information). Serum samples were collected from 50 slaughtered pigs each month. Routine virological surveys were also conducted in Hong Kong in 1977–79 and 1993–94. Swab materials were inoculated into nine-to-ten-day-old embryonated chicken eggs and Madin Darby canine kidney (MDCK) cells; virus isolates were identified and subtyped by haemagglutination inhibition assays as previously described[24].

Virus isolation and sequencing. Viral RNA extraction, complementary DNA synthesis, PCR and sequencing were carried out as described[5,28]. Viral RNA was extracted directly from infected allantoic fluid or cell culture using the QIAamp viral RNA minikit (Qiagen). cDNA was synthesized by reverse transcription; gene amplification by PCR was performed using specific primers for each gene segment. PCR products were purified with the QIAquick PCR purification kit (Qiagen) and sequenced using synthetic oligonucleotides. Reactions were performed using the Big Dye-Terminator v3.1 Cycle Sequencing Reaction Kit on an ABI PRISM 3700 DNA Analyser (Applied Biosystems) following the manufacturer's instructions. All sequences were assembled and edited with Lasergene version 6.1 (DNASTAR). The HA and NA genes were sequenced for all viruses collected in this study and full genome sequencing was conducted for representative viruses, selected on the basis of HA and NA gene diversity and including representative viruses sampled on each positive sampling occasion. All novel reassortants detected on the basis of full genome sequencing were subjected to plaque cloning and full genome sequencing (of at least six randomly selected clones per virus) to confirm that the reassortant was not an artefact of mixed infection.

Antigenic analyses. The antigenic characteristics of SwIV were compared using a haemagglutination inhibition assay with ferret antisera raised against representative influenza A viruses. Ferret antisera raised against Sw/HK/4167/1999 (CS H1N1), Sw/HK/1110/2006 (TRIG H1N2), Sw/HK/NS29/2009 (EA H1N1) and A/California/4/2009 were produced at the Department of Infectious Diseases at St Jude Children's Research Hospital, Memphis, Tennessee and the Department of Microbiology, The University of Hong Kong. The haemagglutination inhibition assay started at 1:40 dilutions for ferret antisera.

To detect antibody prevalence towards major SwIV lineages, we used the haemagglutination inhibition assay with five representative viruses including the antigenically divergent Sw/HK/72/2007-like EA viruses. This allowed us to quantify changes in seroprevalence in serum collected from swine during 2000, 2004, 2009 and 2010.

Experimental infection of pigs. To characterize *in vivo* replicative behaviour of viruses from the major SwIV lineages, we experimentally infected local domestic hybrid (Putian white and Nianbian variant) pigs (*Sus scrofa domesticus*) obtained from a commercial herd and confirmed to be sero-negative and free of influenza virus by HI assays and virus isolation in MDCK cells. Pigs were infected with representative strains belonging to the CS (Sw/HK/4167/1999, Sw/HK/1304/2003), TRIG (Sw/HK/1110/2006), EA (Sw/HK/NS29/2009) and novel EA-reassortant (Sw/HK/72/2007) lineages isolated in this study. Two five-week-old pigs (one male and one female) were intranasally infected with 1 ml of Eagle's minimal essential medium (MEM) containing 10^6 50% tissue culture infectious doses ($TCID_{50}$) of a virus strain. Nasal swabs were collected for 14 d after inoculation from each piglet and placed in 0.6 ml of virus transport medium. Virus shedding in the nasal swabs of pigs was calculated in MDCK by the 50% end-point method[29] and was expressed as $TCID_{50}\,ml^{-1}$ of swab. Animal experiments were carried out in biosafety level three containment facilities at 20–21 °C and 76.5 ± 2.1% relative humidity. Experiments were approved by the Shantou University Medical College and conducted in compliance with university guidelines on animal ethics and welfare.

Molecular evolution and adaptation. Global d_N/d_S rate ratios for each Hong Kong swine lineage and the haemagglutinin gene of European EA viruses were estimated using the codon-based single likelihood ancestor counting method[30]. To determine whether selection was acting differentially on major lineages, the d_N/d_S rate ratio estimate for a lineage was enforced to other co-circulating lineages. A likelihood ratio test was conducted to evaluate whether this fit was significantly worse than unconstrained analysis (and vice versa), with a critical P value of 0.01. This test was repeated using the upper and lower limits of the confidence interval.

Phylogenetic analyses. Phylogenetic trees were inferred using the neighbour-joining method, using genetic distances calculated by maximum likelihood under the Hasegawa, Kishino and Yano (HKY) model with gamma-distributed among-site rate variation (HKY+Γ). The parameters of this model were estimated using maximum likelihood on an initial tree. Temporal phylogenies and rates of evolution were inferred using a 'relaxed molecular clock' model that allows evolutionary rates to vary among lineages in a Bayesian Markov chain Monte Carlo (MCMC) framework[31] This was used to sample phylogenies and dates of divergence while constraining each sequence to its known date of sampling. A model comprising codon-position-specific HKY+Γ substitution models was used. For all analyses employing Bayesian MCMC sampling, a chain length of at least 50 million steps was used with a 10% 'burn-in' removed. At least two independent runs of each chain were performed and compared to ensure adequate sampling. To estimate changes in genetic diversity during our sampling period we used a coalescent-based flexible demographic model[32] to the above MCMC approach. An estimate of the relative genetic diversity ($N_e t$, where N_e is the effective population size and t is the generation time) is obtained by integrating uncertainty across the tree topologies.

28. Poon, L. L. M. *et al.* Rapid detection of reassortment of pandemic influenza H1N1. *Clin. Chem.* **56,** 1340–1344 (2010).
29. Reed, L. J. & Muench, H. A. Simple method of estimating fifty percent endpoints. *Am. J. Hyg.* **27,** 493–497 (1938).
30. Kosakovsky Pond, S. L. & Frost, S. D. W. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* **22,** 1208–1222 (2005).
31. Drummond, A. J., Ho, S. Y., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4,** e88 (2006).
32. Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22,** 1185–1192 (2005).

LETTER

doi:10.1038/nature10003

Profound early control of highly pathogenic SIV by an effector memory T-cell vaccine

Scott G. Hansen[1], Julia C. Ford[1], Matthew S. Lewis[1], Abigail B. Ventura[1], Colette M. Hughes[1], Lia Coyne-Johnson[1], Nathan Whizin[1], Kelli Oswald[2], Rebecca Shoemaker[2], Tonya Swanson[1], Alfred W. Legasse[1], Maria J. Chiuchiolo[3], Christopher L. Parks[3], Michael K. Axthelm[1], Jay A. Nelson[1], Michael A. Jarvis[1], Michael Piatak Jr[2], Jeffrey D. Lifson[2] & Louis J. Picker[1]

The acquired immunodeficiency syndrome (AIDS)-causing lentiviruses human immunodeficiency virus (HIV) and simian immunodeficiency virus (SIV) effectively evade host immunity and, once established, infections with these viruses are only rarely controlled by immunological mechanisms[1–3]. However, the initial establishment of infection in the first few days after mucosal exposure, before viral dissemination and massive replication, may be more vulnerable to immune control[4]. Here we report that SIV vaccines that include rhesus cytomegalovirus (RhCMV) vectors[5] establish indefinitely persistent, high-frequency, SIV-specific effector memory T-cell (T_{EM}) responses at potential sites of SIV replication in rhesus macaques and stringently control highly pathogenic SIV_{MAC239} infection early after mucosal challenge. Thirteen of twenty-four rhesus macaques receiving either RhCMV vectors alone or RhCMV vectors followed by adenovirus 5 (Ad5) vectors (versus 0 of 9 DNA/Ad5-vaccinated rhesus macaques) manifested early complete control of SIV (undetectable plasma virus), and in twelve of these thirteen animals we observed long-term (≥1 year) protection. This was characterized by: occasional blips of plasma viraemia that ultimately waned; predominantly undetectable cell-associated viral load in blood and lymph node mononuclear cells; no depletion of effector-site $CD4^+$ memory T cells; no induction or boosting of SIV Env-specific antibodies; and induction and then loss of T-cell responses to an SIV protein (Vif) not included in the RhCMV vectors. Protection correlated with the magnitude of the peak SIV-specific $CD8^+$ T-cell responses in the vaccine phase, and occurred without anamnestic T-cell responses. Remarkably, long-term RhCMV vector-associated SIV control was insensitive to either $CD8^+$ or $CD4^+$ lymphocyte depletion and, at necropsy, cell-associated SIV was only occasionally measurable at the limit of detection with ultrasensitive assays, observations that indicate the possibility of eventual viral clearance. Thus, persistent vectors such as CMV and their associated T_{EM} responses might significantly contribute to an efficacious HIV/AIDS vaccine.

Conventional prime-boost vaccine regimens with non-persistent vectors lead to lymphoid tissue-based memory T-cell responses ('central memory' or T_{CM}), which deliver peak effector responses only after T_{CM} cells have undergone antigen-stimulated expansion, differentiation and trafficking[6]—too late to effectively control pathogens with the rapid replication and spread kinetics and highly developed immune evasion capabilities of the AIDS-causing lentiviruses[2,4,5]. As T-cell effector responses are likely to be much more effective against the smaller, localized and less diverse viral populations present in the first hours and days of mucosally acquired HIV/SIV infection[2,4,7,8], we proposed that a vaccine able to 'pre-position' differentiated effector cells (T_{EM}) at such early replication sites would demonstrate improved efficacy. Such T_{EM} responses are the hallmark of persistent agents[9,10], prompting our development of SIV vectors based on the persistent β-herpesvirus RhCMV. As recently reported[5] and illustrated in Supplementary Fig. 1, RhCMV/SIV vectors can establish and indefinitely maintain high-frequency SIV-specific, T_{EM}-biased, $CD4^+$ and $CD8^+$ T-cell responses in diverse tissue sites of $RhCMV^+$ rhesus macaques, and in a small efficacy study were associated with early control of intrarectally administered SIV_{MAC239}. To evaluate potential differential effects of persistent vector/T_{EM}-biased versus non-persistent vector/T_{CM}-biased, SIV-specific T-cell responses on the outcome of mucosal SIV_{MAC239} infection, we compared naturally $RhCMV^+$ male rhesus macaques vaccinated with: (1) RhCMV/SIV vectors alone (Group A); (2) RhCMV/SIV vectors followed by replication-defective Ad5 vectors (Group B); and (3) a standard DNA prime/Ad5 vector boost benchmark vaccine (Group C)[11–13] versus unvaccinated control rhesus macaques (Group D; Fig. 1a). RhCMV/SIV vectors efficiently super-infected all Group A and B macaques and elicited robust $CD4^+$ and $CD8^+$ T-cell responses to all vector-encoded SIV proteins (Fig. 1b and Supplementary Figs 2–4). The Ad5 vector boost of Group B macaques, and the DNA/Ad5 regimen given to Group C macaques were also strongly immunogenic (Fig. 1b and Supplementary Figs 3, 4). Although the pattern of development of the SIV-specific T-cell responses differed between these vectors (Supplementary Fig. 3a), the magnitude of the total SIV-specific, $CD4^+$ and $CD8^+$ T-cell responses at the end of the vaccine phase in Groups A, B and C were similar (Fig. 1b and Supplementary Fig. 4). Consistent with previous results[5], RhCMV/SIV-vector-elicited, SIV-specific $CD8^+$ T-cell responses exhibited different epitope targeting than the DNA- and/or Ad5-vector-elicited responses (Supplementary Fig. 3b), and maintained a markedly T_{EM}-biased phenotype over the entire vaccine phase, in contrast to the development of a more T_{CM}-biased response in the DNA/Ad5-vaccinated macaques (Supplementary Fig. 5).

At week 59 after initial vaccination, all rhesus macaques were challenged via the intrarectal route with highly pathogenic SIV_{MAC239} using a repeated, limiting dose protocol[5]. The number of challenges required to achieve measureable infection—plasma viral load > threshold (30 copies ml^{-1})—was not significantly different between Groups A–D (Supplementary Fig. 6), but the subsequent course of infection in these groups was markedly different (Fig. 1c). Of 28 unvaccinated controls (both concurrent and historical), 27 exhibited typical progressive SIV_{MAC239} infection and one exhibited an initially non-progressive infection (transient viraemia) that spontaneously progressed 105 days later. Similarly, all DNA/Ad5-vaccinated macaques (9 of 9) manifested progressive infection, albeit with reduced mean plasma viral load compared to controls (see later). In contrast, 13 of the 24 rhesus macaques that received RhCMV/SIV vectors (6/12 in Group A; 7/12 in Group B) presented with an initial burst of plasma SIV, ranging in magnitude from as few as 60 to as many as 4×10^7 SIV RNA copies ml^{-1}, which was followed by rapid control to undetectable levels (Fig. 1c, d). From 3–18 weeks after infection, all but one of these

[1]Vaccine and Gene Therapy Institute, Departments of Molecular Microbiology and Immunology and Pathology, and the Oregon National Primate Research Center, Oregon Health & Science University, Beaverton, Oregon 97006, USA. [2]AIDS and Cancer Virus Program, SAIC Frederick Inc., National Cancer Institute-Frederick, Frederick, Maryland 21702, USA. [3]International AIDS Vaccine Initiative, Vaccine Design and Development Laboratory, 140 58th Street, Building A, Unit 8J, Brooklyn, New York 11220, USA.

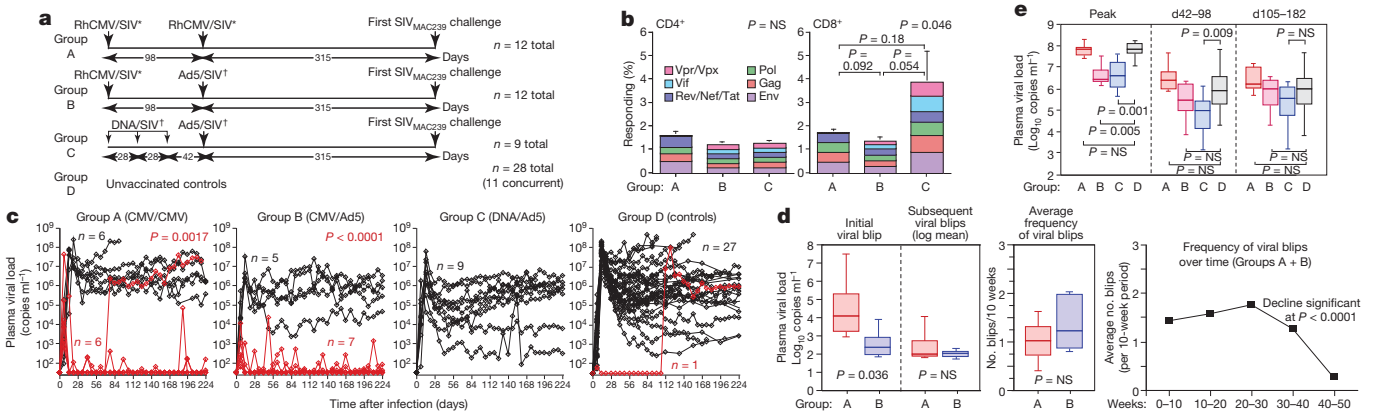


Figure 1 | Immunogenicity and efficacy of RhCMV/SIV vectors. **a**, Schematic of the vaccination protocols used in this study. Note that all rhesus macaques used were naturally RhCMV$^+$ at the start of the study. Asterisk indicates Gag + Rev/Nef/Tat + Env + Pol. Dagger indicates pan-SIV proteome. **b**, Comparison of the mean frequency (± s.e.m.) of the overall SIV-specific CD4$^+$ and CD8$^+$ T-cell responses and the contribution of the designated SIV proteins to these total responses in the blood memory compartments of Groups A–C rhesus macaques at the end of the vaccine phase. The Kruskal–Wallis test was used to determine the significance of differences in total SIV-specific response frequencies among the three vaccine groups, with the Wilcoxon rank sum test used to perform pair-wise analysis for the CD8$^+$ response. As these latter *P* values were >0.05, we concluded that overall response frequencies of Groups A, B and C were not significantly different. NS, not significant. *P* values in top right corners of graphs are for overall Group A versus B versus C. **c**, Outcome of repeated, limiting dose, intrarectal SIV$_{MAC239}$ challenge of Groups A–D. The significance of differences in the fraction of infected rhesus macaques in each group that met controller criteria (see Methods) was determined by Fisher's exact test (closed symbols in Group D are concurrent controls; open, previous controls given the same challenge; red, controllers; black, non-controllers). **d**, Analysis of the magnitude and frequency of plasma viral load 'blips' in Group A and B controllers over the first 50 weeks of infection, with the significance of the differences in blip magnitude and frequency between Groups A and B determined by the Wilcoxon rank sum test, and the significance of the decline in blip frequency of Group A + B macaques after 30 weeks post-infection determined by analysis of variance and linear trend tests. **e**, Comparison of plasma viral loads in Groups A–D rhesus macaques with progressive infection (excluding Group A and B controllers and Group D macaques with protective MHC alleles not represented in Groups A–C) with the significance of differences between Groups A, B and C versus Group D determined by the Wilcoxon rank sum test.

protected rhesus macaques demonstrated one or more repeat episodes of transient viraemia that were always controlled back down to below detection limits (Fig. 1c, d). These periodic viral blips were similar in magnitude in Group A and B controllers, and recurred on average about once every 7 weeks during the first 30 weeks after infection (Fig. 1d). Notably, the frequency of these viral blips declined significantly after week 30 such that by 52 weeks after infection, viral blips were rarely observed (Fig. 1d). No SIV-mediated pathogenesis (loss of effector site CD4$^+$ T cells) was noted in Group A and B controllers (Supplementary Fig. 7), and the vast majority of blood and lymph node mononuclear cell specimens from these macaques were negative for cell-associated SIV RNA and DNA (Supplementary Fig. 8). Six of 12 Group A and 5 of 12 Group B rhesus macaques were not protected in this novel manner, but rather, demonstrated a typical pattern of progressive infection with associated pathogenesis (Fig. 1c, e and Supplementary Fig. 7). The mean peak and plateau phase plasma viral loads of the Group A rhesus macaques with progressive infection were not statistically different from Group D controls (Fig. 1e), indicating that once systemic, progressive infection was established, RhCMV/SIV-vector-elicited responses were unable to control virus replication. The addition of Ad5/SIV vectors in the Group B vaccination regimen was associated with a significantly reduced peak viraemia in Group B macaques with progressive infection compared to Group D controls, but this difference was lost in plateau phase. Consistent with previous reports[11–13], the benchmark DNA/Ad5-vaccinated macaques (Group C) showed significantly reduced log mean peak and early plateau phase (6–14 weeks after infection) plasma viral loads, but for most of these macaques this partial virological control was short-lived, as log mean plasma viral loads in later plateau phase were also not different from Group D controls (Fig. 1c, e). Importantly, the stringent control of SIV infection in protected Group A and B rhesus macaques was not associated with CD8$^+$ T-cell responses restricted by protective MHC alleles (Supplementary Fig. 3b) or with TRIM5 polymorphisms associated with target cell susceptibility to SIV infection (Supplementary Fig. 9).

Taken together, these data indicate that RhCMV/SIV-vector-elicited immune responses mediate a novel pattern of protection in which mucosally administered SIV$_{MAC239}$ is stringently controlled before the onset of progressive, systemic infection. As shown in Fig. 2a and Supplementary Fig. 10, the peak frequencies of SIV-specific CD8$^+$ (but not CD4$^+$) T cells during the vaccine phase (which occurred shortly after the boost), but not the frequencies immediately pre-challenge, significantly correlated with protection in both Groups A and B. These peak responses reflect the level of overall production of SIV-specific CD8$^+$ T cells by the vaccine, and for a T$_{EM}$-biased response would probably parallel the extent of T$_{EM}$ seeding at effector sites. SIV Env-specific antibody responses are not generated by our RhCMV/SIV vectors[5], and did not develop after SIV infection in Group A controllers (Fig. 2b). Although Ad5/SIV-Env-vector-vaccinated rhesus macaques in Group B developed low-titre SIV Env-specific (tissue-culture-adapted SIV$_{MAC251}$-neutralizing) antibody responses before challenge, these titres did not predict control and were not boosted by controlled infection. In contrast, with the exception of rapid progressors, SIV Env-specific antibody responses developed or were boosted in all macaques with systemic, progressive SIV infection. These findings indicate that antibody responses are unlikely to significantly contribute to the protection observed in Group A and B macaques, and further confirm the stringency of protection in RhCMV/SIV-vector-vaccinated controllers, as SIV replication in these macaques produced insufficient antigen to drive humoral immune responses.

We next investigated the effect of SIV infection on the magnitude of the vaccine-elicited T-cell responses. Notably, Group A rhesus macaques showed an almost complete lack of an anamnestic SIV Gag-specific CD4$^+$ or CD8$^+$ T-cell response to either progressive or controlled SIV infection (Fig. 2c and Supplementary Fig. 11). Group B macaques demonstrated a modest anamnestic response in the setting of control, whereas in the setting of progressive infection they manifested a robust anamnestic response, similar to or only slightly less than that observed in Group C macaques. Thus, despite the facts that Group B macaques manifested circulating SIV-specific CD8$^+$ T-cell responses with the characteristic marked T$_{EM}$ bias of RhCMV/SIV-vector-elicited responses (Supplementary Fig. 5), and the early, abrupt RhCMV/SIV-vector-associated pattern of protection (Fig. 1c), these

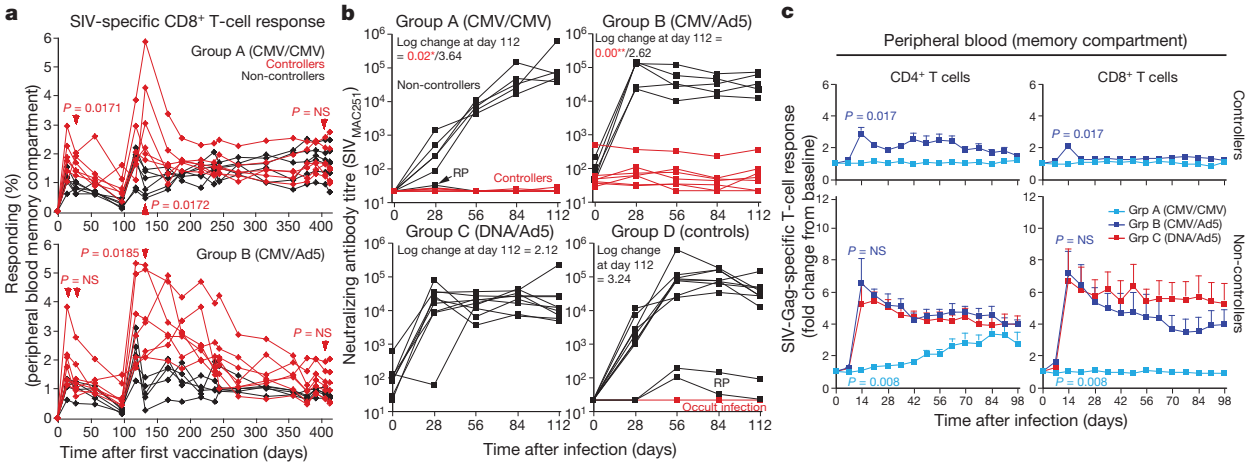


Figure 2 | Immunological correlates of RhCMV/SIV-vector-associated control. **a**, Analysis of total SIV-specific $CD8^+$ T-cell responses (SIV Gag + Rev/Nef/Tat + Pol + Env) in the blood memory compartment during the vaccine phase of Group A and B rhesus macaques with differences in the magnitude of these responses between controllers and non-controllers at the designated time points determined by the Wilcoxon rank sum test. **b**, Comparison of the anti-SIV-Env antibody titres in plasma (as measured by neutralization of tissue-culture-adapted SIV_{MAC251}) before and after infection of controller versus non-controller rhesus macaques among Groups A–C and the concurrent Group D macaques. RP, rapid progressor. Occult infection refers to the initially non-progressive infection (Fig. 1c). The significance of the differences in log change in antibody titre from pre-infection to day 112 post-infection in Group A and B controllers versus Group C macaques was determined by the Wilcoxon rank sum test. $*P < 0.0001$, $**P < 0.005$. **c**, Analysis of the change in the SIV-Gag-specific $CD4^+$ and $CD8^+$ T-cell response frequency after controlled versus progressive infection in Groups A, B and C with the significance of differences in peak response boosting between the designated groups determined by the Wilcoxon rank sum test. Error bars show mean ± s.e.m. *P* values are compared to Group A for the controllers (top panels) and to Group C for the non-controllers (bottom panels).

macaques seemed to maintain a distinct Ad5-vector-elicited, SIV-specific T_{CM} population capable of anamnestic expansion upon either controlled or progressive SIV infection. Importantly, Group A and B controllers robustly responded to infection with *de novo* (Group A) or boosted (Group B) $CD4^+$ and $CD8^+$ T-cell responses to SIV Vif, an antigen not included in the RhCMV/SIV vectors used in this study (Supplementary Fig. 12), confirming both the presence of SIV infection in these macaques, and the normal ability of their naive T-cell (Group A) and T_{CM} (Group B) compartments to respond to the infection. These results indicate that not only does RhCMV/SIV-vector-associated viral control occur in the absence of an overt anamnestic response, but that the SIV-specific T_{EM} populations generated by RhCMV/SIV vectors alone seem unable to significantly expand after infection, regardless of whether antigen levels are limiting (controlled infection) or abundant (progressive infection). This lack of anamnestic expansion may account for the inability of Group A macaques (in contrast to Group B macaques) to manifest any suppression of viral replication once a systemic, progressive infection was established.

The decline in the frequency of SIV RNA blips in the plasma of RhCMV/SIV-vector-vaccinated controllers over time suggests progressive loss of SIV-infected cells, either by immune clearance, virolysis or other attritive mechanisms. To explore the extent of residual infection in long-term RhCMV/SIV-vaccinated controllers, we used monoclonal antibodies to deplete $CD4^+$ or $CD8^+$ lymphocytes from two Group A and two Group B controllers, in comparison to a Group C (DNA/Ad5) and a Group D (unvaccinated) macaque with partial virological control, for each treatment. Administration of the anti-CD4 huOKT4A monoclonal antibody depleted $CD4^+$ T cells, but did not increase plasma viraemia in either Group C and D partial controllers or Group A and B complete controllers (Supplementary Fig. 13). In keeping with previous studies[14,15], $CD8^+$ lymphocyte depletion with the cM-T807 monoclonal antibody did result in a pronounced increase in plasma viral load in Group C and D rhesus macaques with partial control, associated with a robust expansion of SIV Vif-specific $CD4^+$ T cells in effector sites (Fig. 3a). In contrast, $CD8^+$ lymphocyte depletion failed to increase plasma viraemia in RhCMV/SIV-vector-vaccinated controllers,

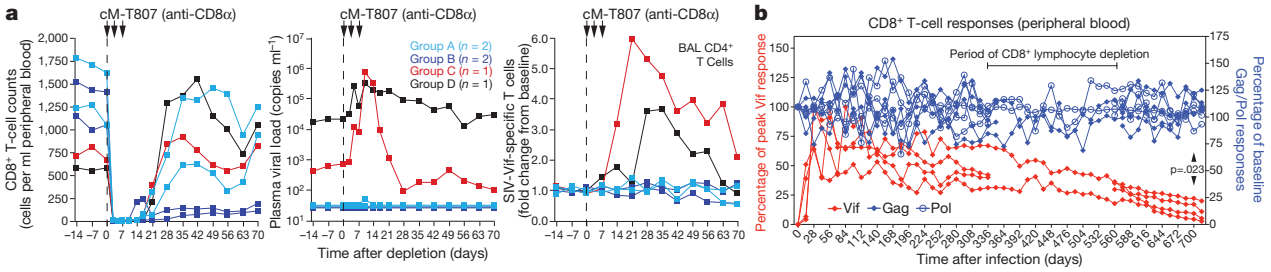


Figure 3 | Immunological characterization of long-term control associated with RhCMV/SIV vector vaccination. **a**, Analysis of the effect of depletion of $CD8^+$ lymphocytes with cM-T807 monoclonal antibody on viral replication and boosting of SIV Vif-specific T-cell responses (in the non-depleted $CD4^+$ subset) in four long-term RhCMV/SIV-vector-vaccinated controllers (two Group A and two Group B rhesus macaques) versus two conventional controllers (one Group C, DNA/Ad5-vaccinated controller; one Group D spontaneous controller). BAL, broncho-alveolar lavage lymphocytes. **b**, Analysis of the frequencies of blood $CD8^+$ T cells specific for SIV proteins that were (Gag, Pol) or were not (Vif) included in the CMV/SIV vectors in the four Group A 'controllers' for which long-term data are available. The response frequencies were normalized to the response frequencies immediately before SIV infection for the Gag- and Pol-specific responses, and to the peak frequencies following SIV infection for the Vif-specific responses. The four rhesus macaques used in this long-term response analysis include those subjected to transient $CD4^+$ or $CD8^+$ lymphocyte depletion (two each). As antigen-specific $CD8^+$ T-cell responses cannot be reliably determined during the period of overall $CD8^+$ lymphocyte depletion, these periods are shown as gaps for two affected rhesus macaques. The significance of differences in the maintenance of response frequencies of Gag- and Pol- versus Vif-specific $CD8^+$ T cells in these rhesus macaques was determined by Wilcoxon rank sum analysis.

and the SIV Vif-specific $CD4^+$ T-cell responses in these macaques were unchanged after depletion, suggesting the absence of even a transient increase in viral replication not detectable by plasma viral load measurements. These studies extend our previous data on the insensitivity of RhCMV/SIV-vector-associated control to $CD8^+$ lymphocyte depletion[5] to rhesus macaques that manifested a higher initial viraemia as well as a period of subsequent, intermittent plasma viral load blips.

As $CD8^+$ T-cell depletion with cM-T807 monoclonal antibody is typically not complete in tissues (Supplementary Fig. 14), lack of viral rebound after such treatment of RhCMV/SIV-vector-vaccinated controllers may simply reflect the potent antiviral function of such residual SIV-specific $CD8^+$ T_{EM} cells or, possibly, the compensatory activity of antiviral $CD4^+$ T_{EM} cells. On the other hand, these observations also raise the possibility that the frequency of SIV-infected and potentially infectious cells in long-term RhCMV/SIV-vector-vaccinated controllers might have been reduced over time to levels that made detectable viral rebound unlikely. In this regard, we found that in Group A controllers, both $CD8^+$ and $CD4^+$ T-cell responses to SIV Vif, an antigen that was not included in the RhCMV/SIV vectors and therefore only available from SIV-infected cells, progressively waned over time to an average of $<$10% of their peak response immediately after (controlled) infection (Fig. 3b and Supplementary Fig. 15). This observation indicates that the numbers of productively infected cells present in these long-term controller macaques are very few, below the threshold necessary to support the initially high-frequency Vif-specific responses. To further examine the extent of residual infection in long-term RhCMV/SIV-vaccinated controllers, we rigorously quantified SIV RNA and DNA at necropsy in four such macaques ($\geq$week 52 after infection; lacking plasma viral load blips for $\geq$10 weeks before necropsy) in comparison to an uninfected macaque, two macaques with SIV infections that were well controlled by standard criteria, and an additional macaque with poorly controlled, progressive SIV infection. As shown in Fig. 4, extensive analysis of lymphoid tissues and immune effector sites of the RhCMV/SIV-vector-vaccinated controllers with ultra-sensitive nested, quantitative reverse transcription polymerase chain reaction (RT–PCR) and polymerase chain reaction (PCR) assays (10 reactions per tissue specimen) demonstrated that cell-associated SIV RNA and DNA were undetectable (0/10 reactions positive) in 72% and 80% of specimens, respectively. In those tissues where viral sequences were detected, the levels were extremely low (approximately one copy per 10^7–10^8 cell equivalents). Notably, the majority of specimens with detectable SIV DNA or RNA (77% and 73%, respectively) were from outside the rectal mucosa. Cell-associated SIV RNA and DNA were not detected in any tissues from an SIV-negative macaque, but were readily detected in all tissues of macaques with conventionally controlled SIV infection. Overall, tissue levels in these conventional controllers averaged $>$3 logs higher than the measureable values of RhCMV/SIV-vector-vaccinated controllers ($P < 0.0001$ by the Wilcoxon rank sum test). Levels of cell-associated SIV were higher still in a macaque with poorly controlled infection. We also assayed lymphoid tissue cells from these macaques for the presence of inducible, replication-competent SIV by co-culture (Supplementary Table 1). All co-cultures (up to 20 replicates per specimen) from RhCMV/SIV-vector-vaccinated controllers were negative for recoverable SIV, whereas replication-competent SIV was readily detected in co-cultures of tissue cells from the conventional controllers. The paucity of SIV nucleic acid and the lack of recoverable SIV in RhCMV/SIV-vector-vaccinated controller macaques are in sharp contrast to the levels of HIV or SIV found in either humans or macaques receiving highly active antiretroviral therapy or in elite controllers[16–21], and suggest an unprecedented level of SIV control and even the possibility of progressive clearance of the SIV infection over time. Importantly, despite little or no SIV replication in the RhCMV/SIV-vector-vaccinated controllers, peripheral blood T cells specific for SIV proteins included in the RhCMV/SIV vectors (for example, Gag and Pol) were stably maintained at high frequency for 700 days after infection ($CD8^+$ T-cell responses with $94 \pm 0.5\%$ T_{EM} phenotype); in

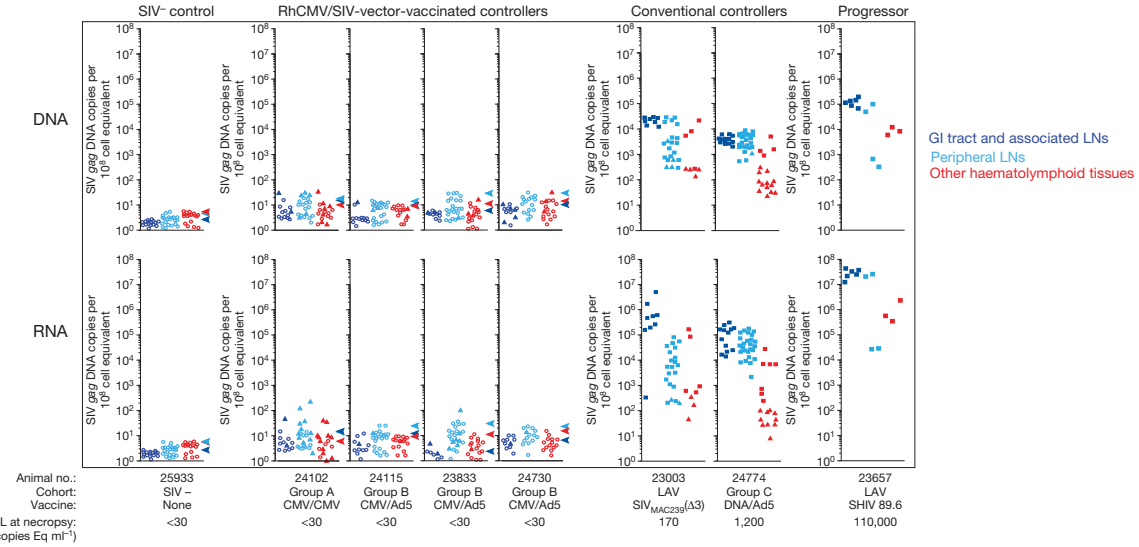


Figure 4 | Measurement of SIV RNA and DNA in long-term RhCMV/SIV-vector-vaccinated controllers. Nested quantitative PCR and quantitative RT–PCR analysis of SIV DNA and RNA, respectively, on tissue obtained at necropsy from an uninfected rhesus macaque, four long-term ($>$52 weeks) RhCMV/SIV-vector-vaccinated controller rhesus macaques (one Group A; three Group B), two conventional controller rhesus macaques (a live attenuated SIV (LAV)-vaccinated macaque that resisted wild-type SIV_{MAC239} challenge 33 and 10 weeks before necropsy, a Group C, DNA/Ad5-vaccine-protected macaque at 55 weeks post-infection), and a rhesus macaque with poorly controlled SIV infection (a LAV-vaccinated macaque 24 weeks after wild-type SIV_{MAC239} challenge). Filled square plot symbols indicate DNA or RNA copy numbers based on directly measured values for samples giving 10/10 replicate reactions positive. Filled triangles indicate results for samples giving at least one, but less than ten, replicate reactions positive, with copy number imputed by Poisson distribution. Open circles indicate specimens that gave 0/10 replicates positive with the symbol's position in the plots at the threshold value corresponding to a Poisson distribution imputed copy number corresponding to 1/10 replicates positive. PVL, plasma viral load. All values are normalized for nucleic acid input. Arrowheads indicate the highest threshold value for negative samples (0/10 replicates positive) for all of the tissues analysed for that macaque. Gastrointestinal (GI) tract and associated lymph nodes (LNs) include colon/rectum, ileum, jejunum, superior/medial/inferior mesenteric and ileocaecal lymph nodes. Peripheral LNs include axillary, submandibular, inguinal, iliosacral and tracheobronchial lymph nodes. Other haematolymphoid tissues include liver, spleen, bone marrow, tonsil and thymus.

contrast to the SIV-infection-elicited Vif-specific responses; Fig. 3b and Supplementary Fig. 15). Thus, persistent RhCMV/SIV vectors provide for long-term maintenance of high-frequency SIV-specific T_{EM} responses, which would otherwise wane with stringent virological control, thereby ensuring continuous, high-level surveillance for SIV-infected cells, even when only rare infected cells are present.

In summary, the 16 long-term RhCMV/SIV-vector-vaccinated controllers described in this and our previous study[5] unequivocally demonstrate a previously undescribed form of immune-mediated control of highly pathogenic SIV in which mucosally acquired infection is arrested before irreversible establishment of disseminated, progressive infection. Although stringently controlled, residual SIV infection is still present for weeks to months in most of these controllers, but wanes over time until eventually it is barely detectable by the most sensitive molecular virological and immunological criteria. The available data strongly indicate that this unique control is related to the high-frequency $CD8^+$, and possibly $CD4^+$, T_{EM}-biased, SIV-specific T-cell responses that are elicited and indefinitely maintained by the persistent RhCMV/SIV vectors, are situated in both mucosal portals of entry and potential sites of distant viral spread, and can protect without anamnestic expansion (see Supplementary Discussion). The ability of RhCMV/SIV vectors to indefinitely maintain SIV-specific T_{EM} responses in these sites, independent of the level of SIV replication, provides for continuous surveillance for SIV-infected cells, preventing relapse and, perhaps, ultimately clearing residual infection. Thus, CMV vectors provide a powerful new approach for HIV/AIDS vaccine development that could be used alone or in combination with complementary vaccine strategies that exploit different HIV immune vulnerabilities.

METHODS SUMMARY

Sixty-seven purpose-bred male rhesus macaques (*Macaca mulatta*) of Indian genetic descent were used in this study. RhCMV/SIV vectors were given subcutaneously at a dose of 5×10^6 plaque-forming units per vector. DNA and Ad5 vectors were given intramuscularly at doses of 1.6 mg per vector and 2×10^{10} particle units per vector, respectively. Rhesus macaques were challenged intrarectally with SIV_{MAC239} using a repeated (weekly) limiting dose protocol[5]. After the onset of infection (plasma viral load ≥30 SIV RNA copy equivalents (Eq) per ml), macaques were followed weekly until onset of AIDS or a minimum of 224 days for progressive infection and 365 days for controlled infection. SIV- and RhCMV-specific $CD4^+$ and $CD8^+$ T-cell responses were measured in mononuclear cell preparations from blood and tissues by flow cytometric intracellular cytokine analysis[5]. SIV Env-specific antibodies were determined by neutralization of tissue-culture-adapted SIV_{MAC251} using a luciferase reporter gene assay[22]. Levels of SIV RNA and DNA in plasma and from isolated cell preparations were quantified by standard quantitative real-time PCR and RT–PCR assays[23,24]. Tissue-associated SIV RNA and DNA at necropsy were quantified by an ultra-sensitive nested, quantitative real-time PCR and RT–PCR approach (see Methods). The presence of inducible, replication-competent SIV in mononuclear cell preparations was detected by co-cultivation with CEMx174 cells, as previously described[16].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 1 December 2010; accepted 17 March 2011.
Published online 11 May 2011.

1. Grovit-Ferbas, K., Pappas, T. & O'Brien, W. A. in *Persistent Viral Infections* (eds Ahmed, R. & Chen, A. I.) 3–45 (John Wiley & Sons, 1999).
2. Haase, A. T. Perils at mucosal front lines for HIV and SIV and their hosts. *Nature Rev. Immunol.* **5,** 783–792 (2005).
3. Goulder, P. J. & Watkins, D. I. Impact of MHC class I diversity on immune control of immunodeficiency virus replication. *Nature Rev. Immunol.* **8,** 619–630 (2008).
4. Haase, A. T. Targeting early infection to prevent HIV-1 mucosal transmission. *Nature* **464,** 217–223 (2010).
5. Hansen, S. G. *et al.* Effector memory T cell responses are associated with protection of rhesus monkeys from mucosal simian immunodeficiency virus challenge. *Nature Med.* **15,** 293–299 (2009).
6. Robinson, H. L. & Amara, R. R. T cell vaccines for microbial infections. *Nature Med.* **11,** S25–S32 (2005).
7. Keele, B. F. *et al.* Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc. Natl Acad. Sci. USA* **105,** 7552–7557 (2008).
8. Chun, T. W. *et al.* Early establishment of a pool of latently infected, resting $CD4^+$ T cells during primary HIV-1 infection. *Proc. Natl Acad. Sci. USA* **95,** 8869–8873 (1998).
9. Bachmann, M. F., Kundig, T. M., Hengartner, H. & Zinkernagel, R. M. Protection against immunopathological consequences of a viral infection by activated but not resting cytotoxic T cells: T cell memory without "memory T cells"? *Proc. Natl Acad. Sci. USA* **94,** 640–645 (1997).
10. Ochsenbein, A. F. *et al.* A comparison of T cell memory against the same antigen induced by virus versus intracellular bacteria. *Proc. Natl Acad. Sci. USA* **96,** 9293–9298 (1999).
11. Casimiro, D. R. *et al.* Attenuation of simian immunodeficiency virus SIVmac239 infection by prophylactic immunization with dna and recombinant adenoviral vaccine vectors expressing Gag. *J. Virol.* **79,** 15547–15555 (2005).
12. Letvin, N. L. *et al.* Preserved $CD4^+$ central memory T cells and survival in vaccinated SIV-challenged monkeys. *Science* **312,** 1530–1533 (2006).
13. Wilson, N. A. *et al.* Vaccine-induced cellular immune responses reduce plasma viral concentrations after repeated low-dose challenge with pathogenic simian immunodeficiency virus SIVmac239. *J. Virol.* **80,** 5875–5885 (2006).
14. Friedrich, T. C. *et al.* Subdominant $CD8^+$ T-cell responses are involved in durable control of AIDS virus replication. *J. Virol.* **81,** 3465–3476 (2007).
15. Reynolds, M. R. *et al.* Macaques vaccinated with live-attenuated SIV control replication of heterologous virus. *J. Exp. Med.* **205,** 2537–2550 (2008).
16. Shen, A. *et al.* Resting $CD4^+$ T lymphocytes but not thymocytes provide a latent viral reservoir in a simian immunodeficiency virus-*Macaca nemestrina* model of human immunodeficiency virus type 1-infected patients on highly active antiretroviral therapy. *J. Virol.* **77,** 4938–4949 (2003).
17. Dinoso, J. B. *et al.* A simian immunodeficiency virus-infected macaque model to study viral reservoirs that persist during highly active antiretroviral therapy. *J. Virol.* **83,** 9247–9257 (2009).
18. North, T. W. *et al.* Viral sanctuaries during highly active antiretroviral therapy in a nonhuman primate model for AIDS. *J. Virol.* **84,** 2913–2922 (2010).
19. Ruiz, L. *et al.* Protease inhibitor-containing regimens compared with nucleoside analogues alone in the suppression of persistent HIV-1 replication in lymphoid tissue. *AIDS* **13,** F1–F8 (1999).
20. Blankson, J. N. *et al.* Isolation and characterization of replication-competent human immuno-deficiency virus type 1 from a subset of elite suppressors. *J. Virol.* **81,** 2508–2518 (2007).
21. Anton, P. A. *et al.* Multiple measures of HIV burden in blood and tissue are correlated with each other but not with clinical parameters in aviremic subjects. *AIDS* **17,** 53–63 (2003).
22. Montefiori, D. C. Evaluating neutralizing antibodies against HIV, SIV, and SHIV in luciferase reporter gene assays. *Curr. Protoc. Immunol.* **12,** Unit 12.11 (2005).
23. Cline, A. N., Bess, J. W., Piatak, M. Jr & Lifson, J. D. Highly sensitive SIV plasma viral load assay: practical considerations, realistic performance expectations, and application to reverse engineering of vaccines for AIDS. *J. Med. Primatol.* **34,** 303–312 (2005).
24. Venneti, S. *et al.* Longitudinal *in vivo* positron emission tomography imaging of infected and activated brain macrophages in a macaque model of human immunodeficiency virus encephalitis correlates with central and peripheral markers of encephalitis and areas of synaptic degeneration. *Am. J. Pathol.* **172,** 1603–1616 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the National Institute of Allergy and Infectious Diseases (RO1 AI060392; contract #HHSN272200900037C); the International AIDS Vaccine Initiative (IAVI) and its donors, particularly the United States Agency for International Development (USAID); the Bill & Melinda Gates Foundation-supported Collaboration for AIDS Vaccine Discovery; the National Center for Research Resources (P51 RR00163; R24 RR016001); and the National Cancer Institute (contract HHSN261200800001E). We thank A. Sylwester, D. Seiss, R. Lum, H. Park and A. Okoye for specialized technical assistance; P. Barry, G. Pavlakis, G. Franchini, C. Miller, N. Wilson, and K. Reimann and Nonhuman Primate Reagent Resource for provision of crucial constructs or reagents; D. Watkins for MHC typing; D. Montefiori for neutralizing antibody assays; N. Letvin and L. Shen for TRIM5a typing; S. Mongoue-Tchokote and M. Mori for statistical assistance; A. Townsend and T. Schroyer for figure preparation; and K. Früh, D. Watkins, B. Beresford, A. McDermott, R. King and W. Koff for discussion and advice.

Author Contributions S.G.H. planned and performed experiments and analysed data, assisted by J.C.F., M.S.L., A.B.V., C.M.H., L.C.-J. and N.W. T.S., A.W.L. and M.K.A. managed the animal protocols. M.A.J. designed, constructed and characterized the RhCMV vectors. M.P. Jr and J.D.L. planned and performed SIV quantification studies, assisted by K.O. and R.S. C.L.P. and M.J.C. designed and constructed the DNA and Ad5 vectors used in this study. J.A.N. was involved in conception of the RhCMV vector strategy. L.J.P. conceived the RhCMV vector strategy, supervised experiments, analysed data and wrote the paper, assisted by S.G.H., M.A.J. and J.D.L.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interest: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to L.J.P. (pickerl@ohsu.edu).

METHODS

Rhesus macaques. A total of 68 purpose-bred juvenile and adult male rhesus macaques (*Macaca mulatta*) of Indian genetic descent were used in the experiments reported in this study, including 61 RhCMV$^+$ macaques in the vaccination/challenge experiment shown in Fig. 1a (Group A and B: $n = 12$ each; Group C: $n = 9$; and Group D: $n = 28$; 11 concurrent and 17 historical controls), 4 long-term RhCMV/Gag-vector-vaccinated macaques, 2 live attenuated SIV-vector-vaccinated/wild-type SIV_{MAC239}-challenged macaques (one controller and one non-controller), and one unvaccinated, uninfected macaque. All macaques were free of cercopithicine herpesvirus 1, D-type simian retrovirus, simian T-lymphotrophic virus type 1 and SIV infection at the start of the study. Group A–C and concurrent Group D included 4 macaques each with the Mamu A*01 allele, but no macaques with the B*08 and B*17 alleles associated with post-peak control of SIV replication[4]. Of the 18 historical Group D rhesus macaques, 11 lacked A*01, B*08 and B*17 alleles, and 6 expressed either B*08 or B*17. The latter macaques were excluded from analysis of viral load in progressive infection, as they expressed protective alleles that were not represented in the vaccine groups. Rhesus macaques were used with approval of the Oregon National Primate Research Center Animal Care and Use Committee under the standards of the US National Institutes of Health Guide for the Care and Use of Laboratory Animals. RhCMV/SIV vectors (RhCMV Gag, Rev/Nef/Tat, Env, Pol1 and Pol2, see later) were given subcutaneously at a dose of 5×10^6 plaque-forming units. Ad5-SIV vectors (Ad5 Gag, Env, Pol, Nef and VVVTR, see later) were given intramuscularly (i.m.) at a dose of 2×10^{10} particle units per vector, and DNA (i.m. at 1.6 mg per vector) was given at weeks 0, 4 and 8 for the DNA prime animals before Ad5-SIV boost at week 14. Plasmids expressing a fusion protein comprised of SIV Vif, Vpr, Vpx, Tat and Rev, or individual Gag, Pol, Env and Nef open reading frames (ORFs) were used for DNA priming. Rhesus macaques were challenged intrarectally with highly pathogenic SIV_{MAC239} using the repeated (weekly), limiting dose protocol described previously[5]. Plasma viral loads were measured weekly, with challenge discontinued the week after detection of >30 copy equivalents (Eq) per ml of SIV RNA (with the challenge preceding the first measured plasma viral load of >30 copy Eq ml^{-1} considered the day of infection). Macaques were considered controllers if plasma viral load became undetectable within 2 weeks of the initial positive plasma viral load and was then maintained below threshold for at least 4 of the subsequent 5 weeks. Challenged macaques were followed until onset of AIDS or a minimum of 224 days for progressive infection and a minimum of 365 days for controlled infection. For CD8$^+$ lymphocyte depletion, rhesus macaques were administered 10, 5, 5 and 5 mg per kg body weight of the humanized monoclonal anti-CD8α antibody cM-T807, on days 0, 3, 7 and 10, respectively[25]. For CD4$^+$ lymphocyte depletion, rhesus macaques were administered one dose of the humanized monoclonal anti-CD4 antibody huOKT4A at 50 mg per kg body weight[26]. Mononuclear cell preparations were obtained from blood, bone marrow, bronchoalveolar lavage (BAL), lymph nodes, spleen, liver, tonsil, thymus and intestinal mucosa as previously described[27–29]. For SIV quantification by nested qPCR/RT–PCR, whole tissue pieces obtained at necropsy were then flash frozen in liquid nitrogen and stored at −80 °C before nucleic acid isolation.

RhCMV/SIV vectors. Construction and characterization of RhCMV Gag, RhCMV Retanef and RhCMV Env has been described[5]. Two additional RhCMV viruses expressing either the 5′ (protease/reverse transcriptase; designated RhCMV Pol1), or 3′ (RNase H/integrase; designated RhCMV Pol2) of SIV_{MAC239} Pol were constructed in an identical fashion by using E/T recombination and the RhCMV (68–1) bacterial artificial chromosome (BAC) (pRhCMV/BAC-Cre)[30]. Deletions were introduced into Pol to inactivate protease (Δ25-DTG-27), reverse transcriptase (Δ184-YMDD-187), RNaseH (ΔE478) and integrase (ΔD64, ΔD116, and ΔE152)[31]. Pol protein expression was placed under control of the EF1α promoter to achieve maximal expression. In all RhCMV/SIV vectors, the SIV antigen-expressing cassettes are inserted into the pRhCMV/BAC-Cre at nucleotide 207,630 within a non-coding region between rh213 and Rh214. RhCMV/SIV viruses were reconstituted by transfection of recombinant BAC DNA into RhCMV-permissive macaque fibroblasts. Following virus reconstitution and BAC cassette 'self-excision', RhCMV/SIV vectors contain the entire wild-type (68–1) RhCMV genome[30]. Vector SIV antigen expression was confirmed by western blot analysis using antibodies to Flag (Sigma-Aldrich; RhCMV Gag), V5 (Invitrogen; RhCMV Retanef), c-Myc/KK45 (Sigma-Aldrich; RhCMV Env), and HA (Sigma-Aldrich; RhCMV Pol1 and Pol2).

DNA vaccines. The plasmid vaccine immunogens used in this study covered the full SIV_{MAC239} (NCBI M33261) genome. ORFs were sequence-optimized for expression in mammalian cells and cloned into a plasmid DNA expression vector. Expression of the intronless coding sequences was controlled by the human CMV (HCMV) promoter/enhancer and the bovine growth hormone polyadenylation signal[32]. Gag (12S), Pol (91S), Env (99S) and Nef (pCMV-Nef) were expressed as single polypeptides, whereas Vif, Vpr, Vpx, Tat and Rev were expressed as a fusion protein (pCMV-VVVTR). The 12S Gag plasmid expresses a myristoylated Gag protein spanning amino acids (aa) 1–508 and lacks two C-terminal aa, but is otherwise similar to a Gag plasmid reported previously[33]. Plasmid 99S expresses a native form of gp160, as previously reported[31,33]. Sequence-optimized Nef without a myristoylation signal was PCR-amplified from 179S plasmid[31], and subsequently transferred into the pCMV vector. The Pol (91S) coding sequence contained deletion mutations to inactivate protease, reverse transcriptase, RNaseH and integrase, as described[31]. Large-scale plasmid production for immunization was prepared by Aldevron, LLC. Expression of plasmids after transient transfection of HEK 293 cells was corroborated by western blot using polyclonal antibodies against SIV_{MAC239} Gag, Env, Nef and Rev proteins made in-house, and anti-SIV_{MAC251} serum from the AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH.

Adenovirus vectors. Plasmid SIV_{MAC239} sequences, again covering the full SIV_{MAC239} genome, were cloned into a human adenovirus serotype 5 (Ad5), which lacks E1A and most of E1B and ΔE3 using the Adeasy Adenoviral vector system (Stratagene) to make the Ad5-SIV Gag, Ad5-SIV Env, Ad5-SIV Pol, AD5-SIV Nef and Ad5-SIV VVVTR vectors. SIV_{MAC239} genes were inserted into the E1 region of the Ad5 under the control of the HCMV immediate early promoter/enhancer and the SV40 polyadenylation signal. All vectors were rescued and propagated in HEK 293 cells and purified by double cesium chloride centrifugation[34]. Dosing was based on the physical number of particles (PU) of Ad5 as measured by spectrophotometry[35]. Expression of SIV proteins from Ad5 vectors after A549 infection was confirmed by western blot as described earlier.

SIV_{MAC239} challenge virus. The pathogenic SIV_{MAC239} challenge virus stock (provided by C. Miller) was generated by expanding the SIV_{MAC239} clone in rhesus macaque peripheral blood mononuclear cells (PBMCs), and was quantified using the sMAGI cell assay and by quantitative RT–PCR for SIV genomic RNA[5].

Immunological assays. SIV- and RhCMV-specific CD4$^+$ and CD8$^+$ T-cell responses were measured in mononuclear cell preparations from blood and tissues by flow cytometric intracellular cytokine analysis, as previously described in detail[5]. Briefly, sequential 15-mer peptides (overlapping by 11 amino acids) comprising the SIV_{MAC239} Gag, Rev/Nef/Tat, Env, Pol, Vif and Vpr/Vpx proteins were used in the presence of co-stimulatory CD28 and CD49d monoclonal antibodies (BD Biosciences). Cells were incubated with antigen and co-stimulatory molecules alone for 1 h, followed by addition of Brefeldin A (Sigma-Aldrich) for an additional 8 h. Co-stimulation without antigen served as a background control. Cells were then stained with fluorochrome-conjugated monoclonal antibodies, flow cytometric data collected on an LSR II (BD Biosciences) and data analysed using the FlowJo software program (version 8.8.6; Tree Star). Response frequencies (CD69$^+$/TNF$^+$ and/or CD69$^+$/IFN-γ$^+$) were first determined in the overall CD4$^+$ and CD8$^+$ population and then memory corrected (with memory T-cell subset populations delineated on the basis of CD28 and CD95 expression), as previously described[5,27]. The data presented as 'end of vaccine phase' response frequencies represent an average of values obtained from samples collected on days 379, 392, 401 and 413 after initial vaccination. Titres of SIV Env-specific antibodies were determined by neutralization of tissue culture-adapted SIV_{MAC251} using a luciferase reporter gene assay[22].

Viral detection assays. Quantitative real-time RT–PCR and PCR assays targeting a highly conserved sequence in Gag were used for standard measurements of plasma SIV RNA and cell-associated SIV RNA and DNA within peripheral blood and lymph node mononuclear cells, as previously described[23,24]. For plasma testing of a sample to score as positive, duplicate amplification reactions yielding ≥30 copy Eq ml^{-1} were required. Isolated viral blips were also repeated in a duplicate sample in almost all instances, and in macaques where infection was manifest only by isolated viral blips (Group A and B controllers), infection was confirmed by the development (Group A) or boosting (Group B) of T-cell responses specific for SIV Vif, an SIV antigen not included in the RhCMV/SIV vectors (Supplementary Fig. 12), as described[5]. To further address the possibility of false positive amplification reactions, we analysed 136 known SIV-negative (pre-challenge) samples from the present study, and from other studies from the same facility and investigators and others. Samples were run in the same laboratory, over approximately the same period of time, using the same procedures, with negative specimens interspersed with positive samples in assay runs. Zero of 136 known SIV-negative samples scored positive by the above criteria, which was significantly different from 84 positive samples (viral 'blips') of 678 total samples (12.4%) obtained from RhCMV/SIV-vector-vaccinated controllers in vaccine Groups A and B ($P < 0.0001$, Fisher's exact test). To more precisely characterize levels of SIV DNA and RNA from tissues of RhCMV/SIV-vector-vaccinated controllers (which were below the sensitivity threshold of the standard assays), we used a new ultra-sensitive, nested, quantitative real-time PCR/RT–PCR method. Tissue pieces were collected directly into extraction tubes and immediately frozen using liquid nitrogen. Samples were stored at −80 °C and handled on dry ice until stabilized in extraction solution. Specimens of approximately 100 mg or less were homogenized in 1 ml of TriReagent (Molecular Research

Center) in 2 ml extraction tubes of Lysing Matrix D using FastPrep instrumentation (MP Biomedicals) according to the manufacturer's recommendations. Total RNA and DNA were prepared from the homogenates following manufacturer's recommendations, but specifically following the alternative, back-extraction method for DNA extraction. Recovered RNA and DNA were dissolved in minimal volumes of 10 mM TrisCl, pH 8.0 and 10 mM TrisCl, pH 9.0, respectively, as appropriate for replicate testing in qRT–PCR and qPCR protocols. Tissue specimens greater than 100 mg in mass were initially pulverized to powder under cryogenic conditions before extraction of RNA and DNA. Pulverization was accomplished in 15 ml polycarbonate extraction tubes with stainless steel grinding balls using rapid vertical shaking, all being maintained at appropriate temperatures in aluminium blocks pre-chilled in liquid N2 (GenoGrinder, SPEX SamplePrep). Pulverized tissue powder was then suspended in 3–10 ml of TriReagent, depending on the starting amount of tissue. Total RNA and DNA were then prepared from 1 ml of TriReagent suspension as described earlier; residual suspension was archived at −80 °C for additional analysis, as necessary. To maximize sensitivity, nested quantitative real-time PCR/RT–PCR protocols were designed to accommodate higher amounts of input nucleic acid, and potential inhibitors, than are typically tolerated in standard assays. Reaction conditions and thermal profiles followed those referenced above for the plasma and isolated cell assays[23,24] with two exceptions: (1) in the quantitative RT–PCR assay, the 'nested' reverse primer, as opposed to random hexamers, was used to prime cDNA synthesis specifically for SIV sequence, thereby avoiding generation of non-specific targets and further enhancing the direct sensitivity of detection of SIV RNA; and (2) 2.5 units of PlatinumTaq (Invitrogen), rather than 1.25 units of TaqGold (Applied Biosystems), were used in the amplification steps. A 'nested' or 'pre-amplification' of cDNA or DNA was performed for 12 cycles with the application of primers, SIVnestF01 (GATTTGGATTAGCAG AAAGCCTGTTG) and SIVnestR01 (GTTG GTCTACTTGTTTTTGGCATAGTTTC), flanking the SIV Gag target region. Five microlitres of this first amplification were then transferred to 50 microlitres of cocktail for amplification of the SIV *gag* DNA target sequence in duplex with amplification of a single copy rhesus CCR5 target sequence for normalization, as referenced earlier[23,24]. Real-time PCR was then performed. For both RNA and DNA determinations, 12 replicate reactions were tested per sample including a spike of 10 copies of RNA or DNA internal control sequence standard in two of the 12 reactions to assess overall amplification efficiency and assess potential inhibition of the PCR or RT–PCR. Samples showing greater than a 5 cycle shift in amplification of the spiked standard, compared to amplification in the absence of specimen nucleic acid, corresponding to less than 74% overall amplification efficiency, were diluted and re-assayed. Quantitative determinations for samples showing amplification in all replicates were derived directly with reference to a standard curve. Quantitative determinations for samples showing less than 10 positive amplifications in replicates were derived from the frequency of positive amplifications, corresponding to the presence of at least one target copy in a reaction, according to a Poisson distribution of a given median copy number per reaction. It should be noted that this assay yielded no positive reactions out of a total of 1,100 total reactions (RNA and DNA) from tissues derived from an SIV-uninfected rhesus macaque, which was significantly different from 178 positive reactions of 4,310 total reactions (4.1%) from tissues derived from the 4 RhCMV/SIV-vector-vaccinated controllers studied at necropsy ($P < 0.0001$; Fisher's exact test). The presence of inducible, replication competent SIV in mononuclear cell preparations derived from different tissue sites at necropsy was detected by co-cultivation with CEMx174 cells, as previously described[16,36]. To detect shedding of RhCMV/SIV vectors in the urine of vaccinated rhesus macaques, virus was concentrated from cleared pan-collected urine and co-cultured with macaque fibroblasts. Cell lysates were collected after development of cytopathic effect or after 28 days, and assessed for vector replication based on expression of SIV antigen-specific epitope tags by western blot analysis[5].

Statistical analysis. We performed statistical analysis with SAS version 9.1 (Statistical Analysis System). Individual tests are described in the figure legends for all analyses. Briefly, Fisher's exact test was used to determine significance of categorical data such as the fraction of controllers versus non-controllers in the different vaccine groups. The Wilcoxon rank sum and Kruskal–Wallis tests were used to compare populations of continuous data for groups of 2 and ≥3, respectively. Analysis of variance and a test for linear trend were used to determine the significance of the reduction in viral blip frequency over time in RhCMV-vector-vaccinated controllers. In all analyses, we used a two-sided significance level (alpha) of 0.05, with correction made for multiple comparisons using the Bonferroni method.

25. Okoye, A. *et al.* Profound $CD4^+/CCR5^+$ T cell expansion is induced by $CD8^+$ lymphocyte depletion but does not account for accelerated SIV pathogenesis. *J. Exp. Med.* **206,** 1575–1588 (2009).
26. Vaccari, M. *et al.* Reduced protection from simian immunodeficiency virus SIVmac251 infection afforded by memory $CD8^+$ T cells induced by vaccination during $CD4^+$ T-cell deficiency. *J. Virol.* **82,** 9629–9638 (2008).
27. Pitcher, C. J. *et al.* Development and homeostasis of T cell memory in rhesus macaque. *J. Immunol.* **168,** 29–43 (2002).
28. Veazey, R. S. *et al.* Gastrointestinal tract as a major site of $CD4^+$ T cell depletion and viral replication in SIV infection. *Science* **280,** 427–431 (1998).
29. Schmitz, J. E. *et al.* Simian immunodeficiency virus (SIV)-specific CTL are present in large numbers in livers of SIV-infected rhesus monkeys. *J. Immunol.* **164,** 6015–6019 (2000).
30. Chang, W. L. & Barry, P. A. Cloning of the full-length rhesus cytomegalovirus genome as an infectious and self-excisable bacterial artificial chromosome for analysis of viral pathogenesis. *J. Virol.* **77,** 5073–5083 (2003).
31. Kulkarni, V. R. *et al.* Comparison of immune responses generated by optimized DNA vaccination against SIV antigens in mice and macaques. *Vaccine* doi:10.1016/j.vaccine.2010.12.056.
32. Rosati, M. *et al.* DNA vaccines expressing different forms of simian immunodeficiency virus antigens decrease viremia upon SIVmac251 challenge. *J. Virol.* **79,** 8480–8492 (2005).
33. Rosati, M. *et al.* DNA vaccination in rhesus macaques induces potent immune responses and decreases acute and chronic viremia after SIVmac251 challenge. *Proc. Natl Acad. Sci. USA* **106,** 15831–15836 (2009).
34. Rosenfeld, M. A. *et al. In vivo* transfer of the human cystic fibrosis transmembrane conductance regulator gene to the airway epithelium. *Cell* **68,** 143–155 (1992).
35. Mittereder, N., March, K. L. & Trapnell, B. C. Evaluation of the concentration and bioactivity of adenovirus vectors for gene therapy. *J. Virol.* **70,** 7498–7509 (1996).
36. Shen, A. *et al.* Novel pathway for induction of latent virus from resting $CD4^+$ T cells in the simian immunodeficiency virus/macaque model of human immunodeficiency virus type 1 latency. *J. Virol.* **81,** 1660–1670 (2007).

LETTER

doi:10.1038/nature09968

Aberrant lipid metabolism disrupts calcium homeostasis causing liver endoplasmic reticulum stress in obesity

Suneng Fu[1], Ling Yang[1], Ping Li[1], Oliver Hofmann[2], Lee Dicker[2], Winston Hide[2], Xihong Lin[2], Steven M. Watkins[3], Alexander R. Ivanov[1] & Gökhan S. Hotamisligil[1,4]

The endoplasmic reticulum (ER) is the main site of protein and lipid synthesis, membrane biogenesis, xenobiotic detoxification and cellular calcium storage, and perturbation of ER homeostasis leads to stress and the activation of the unfolded protein response[1]. Chronic activation of ER stress has been shown to have an important role in the development of insulin resistance and diabetes in obesity[2]. However, the mechanisms that lead to chronic ER stress in a metabolic context in general, and in obesity in particular, are not understood. Here we comparatively examined the proteomic and lipidomic landscape of hepatic ER purified from lean and obese mice to explore the mechanisms of chronic ER stress in obesity. We found suppression of protein but stimulation of lipid synthesis in the obese ER without significant alterations in chaperone content. Alterations in ER fatty acid and lipid composition result in the inhibition of sarco/endoplasmic reticulum calcium ATPase (SERCA) activity and ER stress. Correcting the obesity-induced alteration of ER phospholipid composition or hepatic *Serca* overexpression *in vivo* both reduced chronic ER stress and improved glucose homeostasis. Hence, we established that abnormal lipid and calcium metabolism are important contributors to hepatic ER stress in obesity.

It has been generally accepted that a surplus of nutrients and energy stimulates synthetic pathways and may lead to client overloading in the ER. However, it has not been demonstrated whether increased *de novo* protein synthesis and client loading into the ER and/or a diminished productivity of the ER in protein degradation or folding leads to ER stress in obesity. Intriguingly, dephosphorylation of eukaryotic translation initiation factor 2α (eIF2α) in the liver of high-fat-diet-fed mice reduced the ER stress response[3], indicating that additional mechanisms other than translational upregulation may also contribute to ER dysfunction in obesity. To address these mechanistic questions, we first fractionated ER from lean and obese liver tissues (Supplementary Fig. 1a, b) and then extracted ER proteins for comparative proteomic analysis to examine the status of this organelle in obesity. We identified a total of 2,021 unique proteins (Supplementary Table 1). Among them, 120 proteins were differentially regulated in obese hepatic ER samples (Supplementary Fig. 1c and Supplementary Table 2a, b). We independently validated the differential regulation when possible by immunoblot analyses and verified the fidelity of the system (Supplementary Fig. 1d). Gene ontology analysis identified the enrichment of metabolic enzymes—especially ones involved in lipid metabolism—in the obese ER proteome, whereas protein synthesis and transport functions were overrepresented among downregulated ER proteins (Fig. 1a). Consistently, we found that ER-associated protein synthesis was downregulated in the obese liver as demonstrated by polysome profiling (data not shown), whereas the expression of genes involved in *de novo* lipogenesis (*Fas*, *Scd1*, *Ces1d*, *Dgat2* and *Dak*) and phospholipid synthesis (*Pcyt1a* and *Pemt*) were broadly upregulated (Fig. 1b, c). We also observed upregulation of protein degradation pathways but did not find a broad change in the quantity of ER chaperones (Supplementary Fig. 2 and Supplementary Table 2a). Taken together, these data revealed a fundamental shift in hepatic ER function in obesity from protein to lipid synthesis and metabolism.

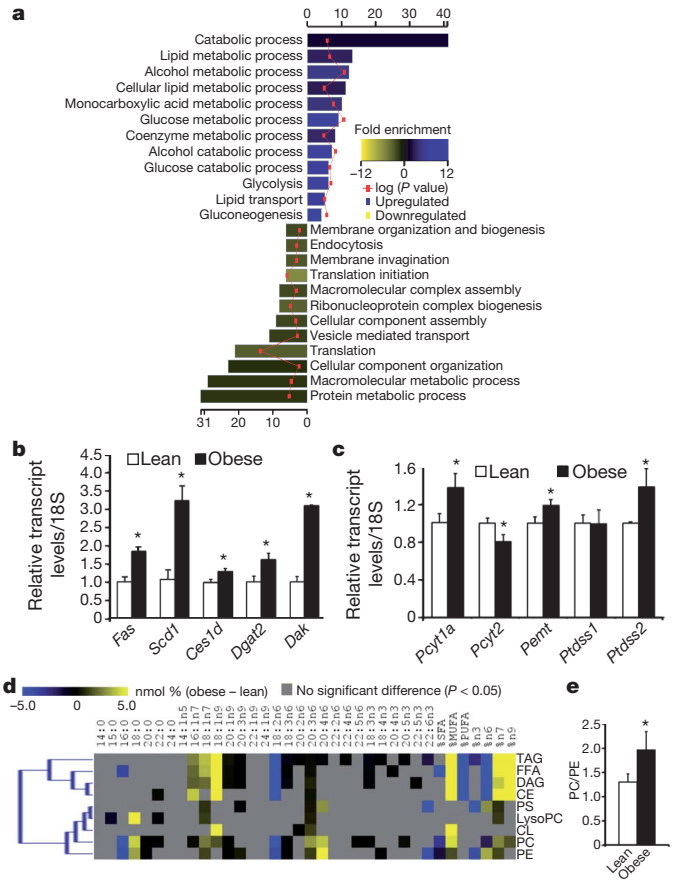


Figure 1 | Proteomic and lipidomic landscape of the lean and obese ER. **a**, Biological pathways associated with significantly regulated proteins in the obese ER proteome. Bar colours indicate the fold enrichment with significance values (negative log of *P* values) superimposed. **b**, **c**, Transcript levels of genes involved in lipid metabolism in the lean and obese mouse liver. **d**, Alterations of liver ER lipidome. Heatmap display of all significant ($P < 0.05$) alterations present between lean and obese ER lipidomes. The colour corresponds to differences in the relative abundance (nmol percentage) of each fatty acid among individual lipid groups detected in the lean and obese liver ER. **e**, The relative abundance of PC and PE in lean and obese liver ER samples. Values are mean ± s.e.m. ($n = 6$ for each group). $^*P < 0.05$, Student's *t*-test.

[1]Departments of Genetics and Complex Diseases, and Nutrition, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [2]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [3]Lipomics Technologies Inc, West Sacramento, California 95691, USA. [4]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA.

The presence of chronic ER stress in obese liver (Supplementary Fig. 2) despite a reduction in ER-associated protein synthesis led us to postulate that the ER stress in obesity may not simply be invoked by protein overloading but also driven by compromised folding capacity, in which lipid metabolism may have a function[4]. For example, the ability of palmitate and cholesterol to induce ER stress in cultured cells correlates with their incorporation into the ER[5,6]. Therefore, we quantitatively determined all major lipid species and their fatty acid composition in ER samples isolated from lean and obese liver along with the diet consumed by these mice (Supplementary Fig. 3 and Supplementary Table 3). First, we found that the fatty acid composition of ER lipids in the lean mouse liver was distinct from corresponding dietary lipids, indicating the contribution of a basal level of *de novo* lipogenesis to the biogenesis of ER membranes *in vivo* (Supplementary Fig. 3a, b and Supplementary Table 3). Almost all ER-derived lipids were composed of significantly higher levels of saturated fatty acids (SFAs) whereas their polyunsaturated fatty acid (PUFA) content was much lower than those of corresponding dietary lipids, indicating that *de novo* synthesized SFAs are preferred over diet-derived PUFAs as the substrate for the synthesis of hepatic ER lipids. Second, the liver ER samples of lean and obese mice also had profoundly different compositions of fatty acids and lipids as illustrated by the clear separation of lean and obese ER lipidome in cluster analysis (Supplementary Fig. 3c). The obese ER was significantly enriched with monounsaturated fatty acids (MUFAs; Fig. 1d), a bona fide product of *de novo* lipogenesis, in liver. Third, the obese ER samples contained a higher level of phosphatidylcholine (PC) as compared to phosphatidylethanolamine (PE) (PC/PE = 1.97 versus 1.3, $P < 0.05$; Fig. 1e and Supplementary Table 3), two of the most abundant phospholipids on the ER membrane. The rise of the PC/PE ratio is probably caused by the upregulation of two key genes involved in PC synthesis and PE to PC conversion: choline-phosphate cytidylyltransferase A (*Pcyt1a*) and phosphatidylethanolamine *N*-methyltransferase (*Pemt*) (Fig. 1c and Supplementary Fig. 3a), and it is consistent with the essential role of PC for lipid packaging in the form of lipid droplets or lipoproteins, both of which are increased in obesity. In contrast, the PC/PE ratio in the lean hepatic ER was essentially identical as it is in the diet (Supplementary Table 3), indicating that the increase of PC/PE ratio in obesity is not due to food consumption, but the result of increased lipid synthesis in the obese liver.

The desaturation of SFAs to MUFAs in the obese liver probably has a protective role in reducing lipotoxicity, whereas the decrease of PUFA content in the ER may limit its reducing capacity and contribute to ER stress[7]. However, a potential role of the PC/PE ratio in regulating ER homeostasis has not been studied before. Previous biochemical studies have shown that increasing PC content in the membrane inhibits the calcium transport activity of SERCA[5,8]. Consistently, we found that the addition of PC to liver-derived microsomes *in vitro* substantially inhibited SERCA activity (Fig. 2a). More importantly, overexpression of the PE to PC conversion enzyme *Pemt* in Hepa1-6 cells significantly inhibited microsomal SERCA activity, indicating that changes in the PC/PE balance in a cellular setting can significantly perturb SERCA function (Fig. 2b, c). As calcium has an important role in mediating chaperone function and protein folding in the ER, and given that SERCA is principally responsible for maintaining calcium homeostasis in this organelle, we postulated that the increased PC/PE ratio in the ER of obese liver might impair ER calcium retention and homeostasis *in vivo*, thereby contributing to protein misfolding and ER stress. In support of this possibility, we found that the calcium transport activity of microsomes prepared from the livers of obese mice was significantly lower than those isolated from lean animals (4.6 ± 0.2 versus 5.3 ± 0.3, $P = 0.046$; Fig. 2d), despite the fact that the SERCA protein level was modestly higher in the former, consistent with an inhibitory role of the PC/PE ratio on SERCA function.

Modest defects in SERCA activity have been implicated in the pathology of Darier's disease[9], and we found that a reduction in SERCA expression *in vivo* (Fig. 2e) and a concurrent reduction in its calcium transport activity (Fig. 2f) potently activated hepatic ER stress in lean mice as evidenced by IRE1α and eIF2α phosphorylation and changes in the expression of GRP78 and GRP94 (Fig. 2g, h). Therefore, there seems to be little redundancy in the function of SERCA beyond physiological fluctuations to maintain ER homeostasis, and the reduction in calcium transport activity could be a potential mechanism of hepatic ER stress in obesity.

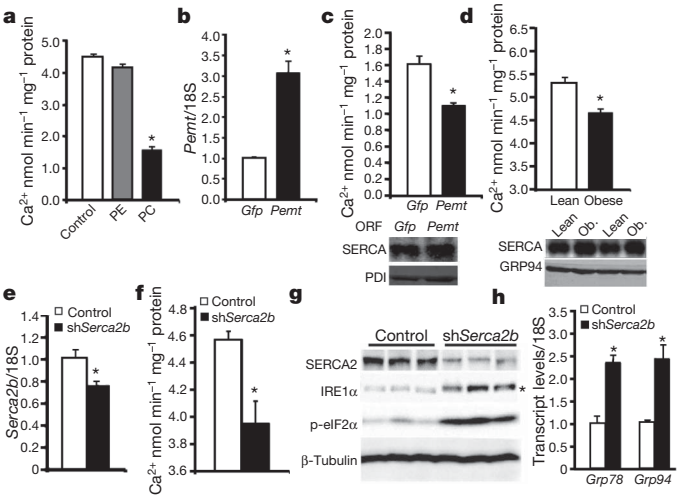


Figure 2 | An increased PC/PE ratio impairs SERCA activity and ER homeostasis. **a**, Calcium transport activity of microsomes loaded with PC and PE *in vitro*. **b**, **c**, Transcript levels of *Pemt* (**b**) and corresponding microsomal calcium transport activities (**c**) of Hepa1-6 cells expressing control (*Gfp*) or mouse *Pemt* open reading frames (ORFs). **d**, Calcium transport activity (top) and SERCA protein levels (bottom) of microsomes prepared from lean and obese mouse liver. **e**–**h**, Liver *Serca2b* transcript levels (**e**) and microsomal calcium transport activities (**f**), immunoblot (**g**) and quantitative RT–PCR (**h**) measurement of ER stress markers in the livers of lean mice expressing either *LacZ* (control) or *Serca2b* shRNAs. Asterisk in **g** denotes the phosphorylated IRE1α and in other panels denotes significant difference (*$P < 0.05$, $n = 4$) by Student's *t*-test. Values are mean ± s.e.m.

We carried out two different but complementary approaches to correct aberrant lipid metabolism induced SERCA dysfunction and examined the effects on ER homeostasis in the obese liver. If the alteration in PC/PE ratio seen in obese liver is a significant contributor to ER stress, correction of this ratio to lean levels by reducing *Pemt* expression should improve calcium transport defects and produce beneficial effects on hepatic ER stress and metabolism. Using an adenovirally expressed short hairpin RNA (shRNA), we were able to achieve ~50–70% suppression of the *Pemt* transcript in obese liver (Supplementary Fig. 4a). As postulated, suppression of *Pemt* led to a decrease of PC content from ~39% to ~33%, which was compensated by an ~7% increase of PE content from ~17% to 24% (Supplementary Table 4). As a result, the PC/PE ratio is reduced to 1.3 (equivalent to the lean ratio), as compared to 2.0 detected in the ER of the obese liver (Fig. 3a). The reduction of the PC/PE ratio was accompanied by a significant improvement in the calcium transport activity of the ER prepared from the *Pemt*-knockdown obese mice (Fig. 3b). As the improvement of calcium transport function occurred with few and minor changes in the overall fatty acid composition of ER (Supplementary Fig. 4b, c and Supplementary Table 5), our results confirmed the rise in PC/PE ratio as an inhibitory factor of SERCA activity in obesity. More importantly, hepatic ER stress indicators including the phosphorylation of IRE1α and eIF2α as well as the expression of C/EBP homologous protein (CHOP), homocysteine-inducible, ER stress-inducible protein (HERP) and Der1-like domain family member 2 (DERL2) were all reduced upon suppression of *Pemt* in obese mice (Fig. 3c, d and Supplementary Fig. 4d). Relief of chronic ER stress in leptin-deficient ($Lep^{-/-}$) mice has been associated with improvement of hepatic steatosis and glucose homeostasis[10,11].

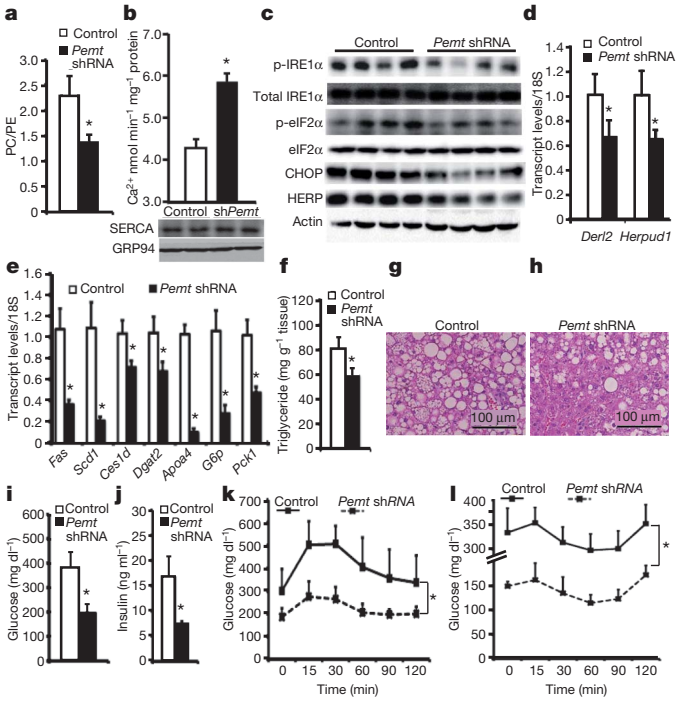


Figure 3 | Suppression of liver *Pemt* expression corrects the ER PC/PE ratio, relieves ER stress and improves systemic glucose homeostasis in obesity. **a**, **b**, PC/PE ratio (**a**) and calcium transport activity (**b**) of liver ER from $Lep^{-/-}$ mice expressing *LacZ* (control) or *Pemt* shRNAs. **c**, **d**, Immunoblot (**c**) and quantitative PCR (**d**) measurement of ER stress markers in the liver. **e**–**h**, Expression of hepatic lipogenesis and gluconeogenesis genes (**e**), triglyceride content (**f**) and haematoxylin & eosin staining (**g** and **h**) of liver samples. **i**, **j**, Plasma glucose (**i**) and insulin (**j**) levels in control and *Pemt* shRNA-treated $Lep^{-/-}$ mice after 6 h food withdrawal. **k**, **l**, Plasma glucose levels of control and *Pemt* shRNA-treated $Lep^{-/-}$ mice after intraperitoneal administration of either 1 g kg^{-1} of glucose (**k**) or 1 IU kg^{-1} of insulin (**l**). All data are mean ± s.e.m. ($n = 4$ for **a**–**e**, $n = 6$ for **f**–**l**), $^*P < 0.05$ (one-way ANOVA for data presented in **k** and **l**, and Student's *t*-test for others).

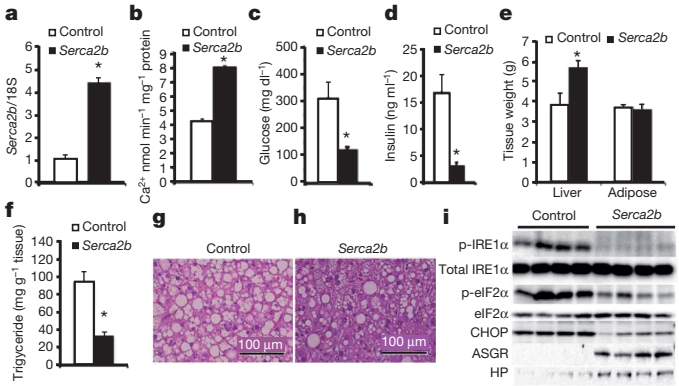


Figure 4 | Exogenous *Serca* expression alleviates ER stress and improves systemic glucose homeostasis. **a**, **b**, Liver *Serca2b* transcript levels (**a**) and microsomal calcium transport activities (**b**) of control or *Serca2b*-overexpressing obese mice. **c**–**e**, Plasma glucose (**c**), plasma insulin levels (**d**) and tissue weights (**e**) of $Lep^{-/-}$ mice as in panel **a**. **f**–**i**, Triglyceride content (**f**), haematoxylin & eosin staining (**g**, **h**) and immunoblot analyses (**i**) of ER stress markers (IRE1α and eIF2α phosphorylation, and CHOP) and secretory proteins (ASGR and HP) in the obese liver expressing *Serca2b* compared to controls. All values are mean ± s.e.m. ($n = 4$ for **a** and **b**, $n = 6$ for **c**–**h**). $^*P < 0.05$ (Student's *t*-test).

Consistently, genes involved in hepatic lipogenesis (*Fas*, *Scd1*, *Ces1d*, *Dgat2*) and lipoprotein synthesis (*Apoa4*) were significantly downregulated in the obese liver after suppression of *Pemt* (Fig. 3e). As a result, these mice exhibited a significant reduction in hepatic steatosis and liver triglyceride content (Fig. 3f–h). Genes involved in glucose production (*G6pc*, *Pck1*) in the liver were significantly downregulated (Fig. 3e), and there were also significant reductions in both hyperglycaemia and hyperinsulinaemia in obese mice after the suppression of hepatic *Pemt* expression (Fig. 3i, j). Glucose and insulin tolerance tests revealed significantly enhanced glucose disposal after *Pemt* suppression (Fig. 3k, l). A similar phenotype is also observed upon suppression of hepatic *Pemt* in high-fat-diet-induced obesity, with reduced ER stress and improved glucose homeostasis (Supplementary Fig. 5). These data are consistent with the phenotype seen in *Pemt*-deficient mice, which exhibit protection against diet-induced insulin resistance and atherosclerosis[12]. Therefore, correcting the PC/PE ratio of the ER can significantly improve calcium transport defects, reduce ER stress and improve metabolism, supporting the hypothesis that changes in lipid metabolism contribute to SERCA dysfunction, ER stress and hyperglycaemia in both genetic- and diet-induced models of obesity.

We then carried out overexpression of hepatic *Serca in vivo* to overcome the partial inhibition of SERCA activity by PC (Fig. 4a). Indeed, exogenous SERCA expression in the liver of $Lep^{-/-}$ mice improved the calcium import activity of the ER (Fig. 4b), restored euglycaemia and normoinsulinaemia within a few days, and markedly improved glucose tolerance (Fig. 4c, d and Supplementary Fig. 6). Upon *Serca* expression, the liver showed an increase in size but a marked reduction of lipid infiltration (Fig. 4e–h) and suppression of IRE1α and eIF2α phosphorylation, along with a significant reduction in CHOP levels (Fig. 4i). In these liver samples, there was also a marked increase in two secretory proteins that were otherwise diminished in obesity: asialoglycoprotein receptor (ASGR) and haptoglobin (HP) (Fig. 4i). As the folding and maturation of ASGR is most sensitive to perturbations of calcium homeostasis in the ER[13], our results indicate that exogenously increased SERCA expression restored calcium homeostasis and relieved at least some aspects of chronic ER stress in the obese liver. Taken together, these data reinforce the hypothesis that lipid-driven alterations and ER calcium homeostasis are important contributors to hepatic ER stress in obesity.

The chronic activation of ER stress markers has been observed in a variety of experimental obese models as well as in obese humans[14]. Furthermore, treatment of obese mice and humans with chemical chaperones results in increased insulin sensitivity[10,15]. Our systematic, compositional and functional characterization of hepatic ER landscape from lean and obese mice revealed a diametrically opposite regulation of ER functions regarding protein and lipid metabolism and revealed mechanisms giving rise to ER stress. In particular, an increase in the PC/PE ratio in the ER, driven by the upregulation of *de novo* lipogenesis in obesity, was linked to SERCA dysfunction and chronic ER stress *in vivo*. During the review of this manuscript, a study reported downregulation of the SERCA protein level in obese liver[16], which was not evident in our analysis and seemed to have resulted from the choice of methodology in ER protein preparations (Supplementary Fig. 7). Nevertheless, other mechanisms such as oxidative and inflammatory changes associated with obesity can also perturb ER homeostasis by affecting ER calcium fluxes[17–19] and will be important to study in the future.

The identification of a lipid-driven calcium transport dysfunction and ER stress provides a fundamental framework for understanding the pathogenesis of hepatic lipid metabolism and chronic ER stress in obesity. First, excessive food intake inevitably stimulates lipogenesis for energy storage, and PC is the preferred phospholipid coat of lipid droplets and lipoproteins[20]. Therefore, there is a biological need for the synthesis of more PC for packaging and storing the products of hepatic lipogenesis. Second, *de novo* fatty acid synthesis in the obese liver produces ample amounts of MUFA, which is effectively incorporated into PC but not PE, which further distorts the PC/PE ratio and impairs ER function. The resulting ER stress facilitates the secretion of excessive lipids from the liver without ameliorating hyperinsulinaemia-induced lipogenesis[21], and thus hepatosteatosis and ER stress ensue. As a result,

relieving ER stress in obesity may ultimately depend on breaking this 'lipogenesis–ER-stress–lipogenesis' vicious cycle and restoring ER folding capacity. Therefore, we suggest that genetic, chemical or dietary interventions that modulate hepatic phospholipid synthesis and/or ER calcium homeostasis function might represent a new set of therapeutic opportunities for common chronic diseases associated with ER stress, such as obesity, insulin resistance and type 2 diabetes.

METHODS SUMMARY

Male leptin-deficient ($Lep^{-/-}$) and wild-type littermates in the C57BL/6J background were either bred in-house or purchased from the Jackson Laboratory (strain B6.V-Lep^{ob}/J, stock number 000632). Transduction of adenoviruses (serotype 5, Ad5) for the expression of open reading frames (ORFs) or shRNAs was carried out between 10–11 weeks after birth, and all mice were killed between 12–13 weeks of age, unless noted otherwise. ER fractionation for proteomic and lipidomic analysis were carried out as previously described[22]. Calcium transport experiments were performed as previously described[23], with some modifications. Quantitative RT–PCR, western blot analysis, histology and *in vivo* animal experiments were carried out as previously described[10,24]. Oligonucleotide sequences used in this study are listed in Supplementary Table 6. Detailed experimental procedures and protocols are described in the Supplementary Material.

Received 13 October 2010; accepted 22 February 2011.
Published online 1 May 2011.

1. Ron, D. & Walter, P. Signal integration in the endoplasmic reticulum unfolded protein response. *Nature Rev. Mol. Cell Biol.* **8,** 519–529 (2007).
2. Hotamisligil, G. S. Endoplasmic reticulum stress and the inflammatory basis of metabolic disease. *Cell* **140,** 900–917 (2010).
3. Oyadomari, S. *et al.* Dephosphorylation of translation initiation factor 2α enhances glucose tolerance and attenuates hepatosteatosis in mice. *Cell Metab.* **7,** 520–532 (2008).
4. Erbay, E. *et al.* Reducing endoplasmic reticulum stress through a macrophage lipid chaperone alleviates atherosclerosis. *Nature Med.* **15,** 1383–1391 (2009).
5. Li, Y. *et al.* Enrichment of endoplasmic reticulum with cholesterol inhibits sarcoplasmic-endoplasmic reticulum calcium ATPase-2b activity in parallel with increased order of membrane lipids: implications for depletion of endoplasmic reticulum calcium stores and apoptosis in cholesterol-loaded macrophages. *J. Biol. Chem.* **279,** 37030–37039 (2004).
6. Borradaile, N. M. *et al.* Disruption of endoplasmic reticulum structure and integrity in lipotoxic cell death. *J. Lipid Res.* **47,** 2726–2737 (2006).
7. Kim, S. J. *et al.* Omega-3 and omega-6 fatty acids suppress ER- and oxidative stress in cultured neurons and neuronal progenitor cells from mice lacking PPT1. *Neurosci. Lett.* **479,** 292–296 (2010).
8. Cheng, K. H., Lepock, J. R., Hui, S. W. & Yeagle, P. L. The role of cholesterol in the activity of reconstituted Ca-ATPase vesicles containing unsaturated phosphatidylethanolamine. *J. Biol. Chem.* **261,** 5081–5087 (1986).
9. Miyauchi, Y. *et al.* Comprehensive analysis of expression and function of 51 sarco(endo)plasmic reticulum Ca^{2+}-ATPase mutants associated with Darier disease. *J. Biol. Chem.* **281,** 22882–22895 (2006).
10. Ozcan, U. *et al.* Chemical chaperones reduce ER stress and restore glucose homeostasis in a mouse model of type 2 diabetes. *Science* **313,** 1137–1140 (2006).
11. Kammoun, H. L. *et al.* GRP78 expression inhibits insulin and ER stress-induced SREBP-1c activation and reduces hepatic steatosis in mice. *J. Clin. Invest.* **119,** 1201–1215 (2009).
12. Jacobs, R. L. *et al.* Impaired *de novo* choline synthesis explains why phosphatidylethanolamine *N*-methyltransferase-deficient mice are protected from diet-induced obesity. *J. Biol. Chem.* **285,** 22403–22413 (2010).
13. Lodish, H. F. & Kong, N. Perturbation of cellular calcium blocks exit of secretory proteins from the rough endoplasmic reticulum. *J. Biol. Chem.* **265,** 10893–10899 (1990).
14. Gregor, M. F. *et al.* Endoplasmic reticulum stress is reduced in tissues of obese subjects after weight loss. *Diabetes* **58,** 693–700 (2009).
15. Kars, M. *et al.* Tauroursodeoxycholic acid may improve liver and muscle but not adipose tissue insulin sensitivity in obese men and women. *Diabetes* **59,** 1899–1905 (2010).
16. Park, S. W., Zhou, Y., Lee, J. & Ozcan, U. Sarco(endo)plasmic reticulum Ca^{2+}-ATPase 2b is a major regulator of endoplasmic reticulum stress and glucose homeostasis in obesity. *Proc. Natl Acad. Sci. USA* **107,** 19320–19325 (2010).
17. Li, S. Y. *et al.* Cardiac contractile dysfunction in Lep/Lep obesity is accompanied by NADPH oxidase activation, oxidative modification of sarco(endo)plasmic reticulum Ca^{2+}-ATPase and myosin heavy chain isozyme switch. *Diabetologia* **49,** 1434–1446 (2006).
18. Cardozo, A. K. *et al.* Cytokines downregulate the sarcoendoplasmic reticulum pump Ca^{2+} ATPase 2b and deplete endoplasmic reticulum Ca^{2+}, leading to induction of endoplasmic reticulum stress in pancreatic β-cells. *Diabetes* **54,** 452–461 (2005).
19. Li, G. *et al.* Role of ERO1-α-mediated stimulation of inositol 1,4,5-triphosphate receptor activity in endoplasmic reticulum stress-induced apoptosis. *J. Cell Biol.* **186,** 783–792 (2009).
20. Schiller, J. *et al.* Lipid analysis of human HDL and LDL by MALDI-TOF mass spectrometry and (31)P-NMR. *J. Lipid Res.* **42,** 1501–1508 (2001).
21. Brown, M. S. & Goldstein, J. L. Selective versus total insulin resistance: a pathogenic paradox. *Cell Metab.* **7,** 95–96 (2008).
22. Cox, B. & Emili, A. Tissue subcellular fractionation and protein extraction for use in mass-spectrometry-based proteomics. *Nature Protocols* **1,** 1872–1878 (2006).
23. Moore, L., Chen, T., Knapp, H. R. Jr & Landon, E. J. Energy-dependent calcium sequestration activity in rat liver microsomes. *J. Biol. Chem.* **250,** 4562–4568 (1975).
24. Cao, H. *et al.* Identification of a lipokine, a lipid hormone linking adipose tissue to systemic metabolism. *Cell* **134,** 933–944 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Porter, E. Freeman and R. Davis for technical assistance. The anti-HERP antibody is a gift of Y. Hirabayashi. We thank the members of the G.S.H. laboratory for scientific discussions and critical reading of the manuscript. This work was supported in part by the National Institutes of Health (DK52539 and 1RC4-DK090942) and a research grant from Syndexa Pharmaceuticals to G.S.H. S.F. was supported in part by the NIH/NIEHS postdoctoral training grant (T32ES007155).

Author Contributions S.F. designed, performed experiments, analysed and interpreted the results and wrote the manuscript; L.Y. and P.L. performed some animal experiments; O.H., L.D., W.H. and X.L. performed statistical and bioinformatic analysis of the proteomic data; S.W.M quantified the lipid composition of ER and analysed the data; A.R.I. analysed the protein composition of ER; G.S.H. generated the hypothesis, designed the project, analysed and interpreted the data and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to G.S.H. (ghotamis@hsph.harvard.edu).

LETTER

doi:10.1038/nature09970

Neural crest regulates myogenesis through the transient activation of NOTCH

Anne C. Rios[1], Olivier Serralbo[1]*, David Salgado[1]* & Christophe Marcelle[1]

How dynamic signalling and extensive tissue rearrangements interact to generate complex patterns and shapes during embryogenesis is poorly understood[1–3]. Here we characterize the signalling events taking place during early morphogenesis of chick skeletal muscles. We show that muscle progenitors present in somites require the transient activation of NOTCH signalling to undergo terminal differentiation. The NOTCH ligand Delta1 is expressed in a mosaic pattern in neural crest cells that migrate past the somites. Gain and loss of Delta1 function in neural crest modifies NOTCH signalling in somites, which results in delayed or premature myogenesis. Our results indicate that the neural crest regulates early muscle formation by a unique mechanism that relies on the migration of Delta1-expressing neural crest cells to trigger the transient activation of NOTCH signalling in selected muscle progenitors. This dynamic signalling guarantees a balanced and progressive differentiation of the muscle progenitor pool.

Early skeletal muscle (the primary myotome, which is composed of mononucleated post-mitotic muscle fibres, or myocytes) is formed from the generation of muscle cells at the four borders of the dermomyotome, the dorsal-most epithelial compartment of somites[4–10]. Most of the dermomyotome undergoes an epithelial to mesenchymal transition that leads to the emergence of a population of resident muscle progenitors that massively contributes to the growth of all trunk muscles[11–14]. The medial border of the dermomyotome, the dorsomedial lip (DML), remains epithelial for a considerable period of time, during which it generates muscle cells that contribute to the growth of the primary myotome. DML stem/progenitor cells can adopt two fates during the first days of embryonic muscle development[4–6]: to self-renew and remain in the epithelial border of the dermomyotome or to translocate in the myotome and undergo terminal myogenic differentiation. How this balance is regulated is unknown.

In the chick embryo, the epithelial DML population comprises a majority (77%) of PAX7-positive cells interspersed by (23%) PAX7/MYF5-positive cells (Supplementary Figs 1a–e, 2a–e). In the transition zone, cells shut-off the expression of PAX7, but maintain MYF5 expression. Fully elongated myocytes express skeletal muscle myosin heavy chain (MyHC; also known as MYC). NOTCH family members are expressed in the DML, the transition zone and the myotome during the first phase of myogenesis (Fig. 1a)[15]. The NOTCH target genes *HES1*/cHairy2 and lunatic fringe (*LFNG*) are expressed in a salt-and-pepper pattern within the DML. Many transition zone cells express *HES1*, whereas *LFNG* expression is low in this region. Their expression is low in the myotome (Fig. 1a, b and Supplementary Fig. 2f). Both genes act as bona fide NOTCH targets in somites, as their messenger RNA expression is upregulated after electroporation of a constitutive form of NOTCH (NOTCH intracellular domain (NICD); Supplementary Fig. 2g–i, m–o). To quantify the distribution of NOTCH activity, we electroporated a NOTCH reporter construct consisting of the mouse *Hes1* promoter region upstream of a destabilized GFP (d2EGFP; half life, 2 h), that efficiently responds to NOTCH activation and inhibition (Supplementary Fig. 2j–l). We co-electroporated a human histone H2B (H2B)–RFP reporter gene driven by an ubiquitous promoter to evaluate the normal distribution of electroporated cells. After 24 h, 11% of H2B–RFP-positive cells were HES1–d2EGFP-positive (Fig. 1c, d and Supplementary Fig. 2l). The H2B–RFP-positive cells were distributed among PAX7- (60%), MYF5- (46%) and MyHC-positive (10%) populations. In contrast, nearly all (92%) HES1-positive cells were MYF5-positive (distributed in the DML and the transition zone), whereas only 24% and 2% expressed PAX7 and MyHC, respectively (Fig. 1e–i and Supplementary Fig. 3a, b). We followed the morphogenic movements of NOTCH-activating cells using live video microscopy. Epithelial cells that activated the NOTCH reporter in the DML rapidly translocated in the transition zone (Fig. 1j–l and

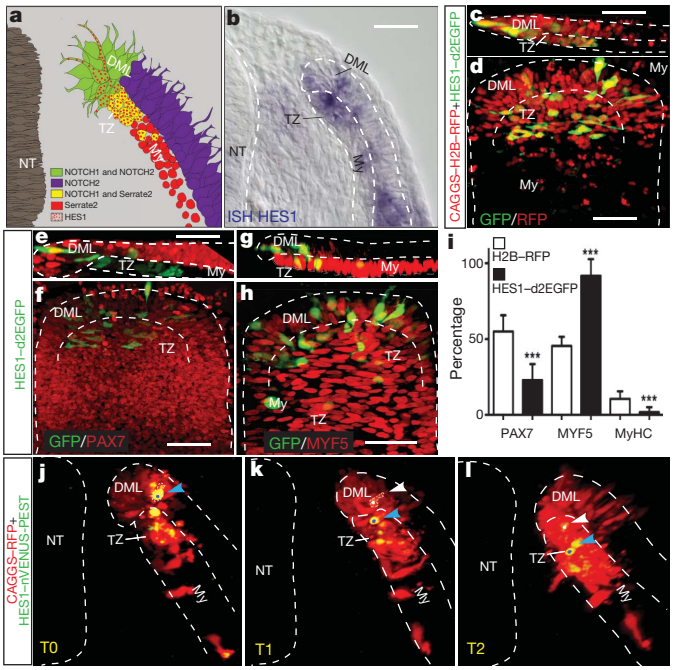


Figure 1 | Notch is active during early myogenesis. **a**, Scheme showing the expression of NOTCH signalling family members in the DML, transition zone (TZ) and myotome. Serrate2 is also known as JAGGED2. **b**, Expression of chick *HES1*/chick Hairy2 in the DML and the TZ. ISH, *in situ* hybridization. **c**–**h**, Confocal stacks showing the expression (in green) of a HES1–d2EGFP reporter construct and (in red) RFP (**c**, **d**), PAX7 (**e**, **f**) and MYF5 (**g**, **h**) in dorsal (**d**, **f**, **h**) and transverse (**c**, **e**, **g**) views of somites 24 h after electroporation. **i**, Quantification of **c**–**h**. Error bars show standard deviation (s.d.). ***$P < 0.0001$. **j**–**l**, Time-lapse confocal analysis (see Supplementary Movie 1) showing the translocation of two NOTCH-activating DML cells (blue and white arrowheads) into the transition zone. T0, start of the movie; T1, 4 h 50 min after the start; T2, 10 h after the start. My, myotome; NT, neural tube. Scale bars, 50 μm.

[1]EMBL Australia; Australian Regenerative Medicine Institute (ARMI), Monash University, Building 75, Clayton, Victoria 3800, Australia.
*These authors contributed equally to this work.

Supplementary Movie 1). We followed their fate as they further differentiated, using a construct that contains the *Hes1* promoter region upstream of a stable GFP (EGFP half life of over 24 h). After 24 and 48 h, the proportion of MyHC-positive myocytes was more than twice (24 h: 34%; 48 h: 60%) that of control RFP-electroporated embryos (24 h: 13%; 48 h: 28%; Supplementary Fig. 3c–g), further indicating that activation of NOTCH signalling is associated with myogenesis.

Altogether, this indicates that NOTCH signalling is activated in DML cells that engage in the myogenic program before they translocate into the transition zone. NOTCH signalling remains active in the transition zone and is extinguished before cells undergo terminal myogenic differentiation and elongate into myocytes.

We inhibited NOTCH activity in somites using a truncated, dominant-negative form of the NOTCH transcriptional co-activator mastermind (DN MAML1)[16,17] and small interfering RNAs (siRNAs) against *NOTCH1* (ref. 18). DN MAML1 and the *NOTCH1* siRNA gave similar results one day later, that is, a drastic reduction of myogenic differentiation (Fig. 2). This was characterized by a sharp reduction of MYF5-positive cells (7% DN MAML1 ; 3% siRNA *NOTCH*), compared to controls (49% CAGGS–IRES–EGFP ; 53% siRNA luciferase), and by a halt of terminal differentiation (0% MyHC-positive cells for DN MAML1 and siRNA *NOTCH*; controls: 12% CAGGS–IRES–EGFP; 8% siRNA luciferase; Fig. 2s, t), with no change in dermomyotome cell proliferation (Supplementary Fig. 4a–d). Virtually all cells in which NOTCH signalling was inhibited remained epithelial in the dermomyotome (98% DN MAML1; 97% siRNA *NOTCH*; controls: 56% CAGGS–IRES–EGFP; 57% siRNA luciferase). Altogether, this strongly indicates that NOTCH signalling is necessary for the initial phases of myogenesis of DML cells.

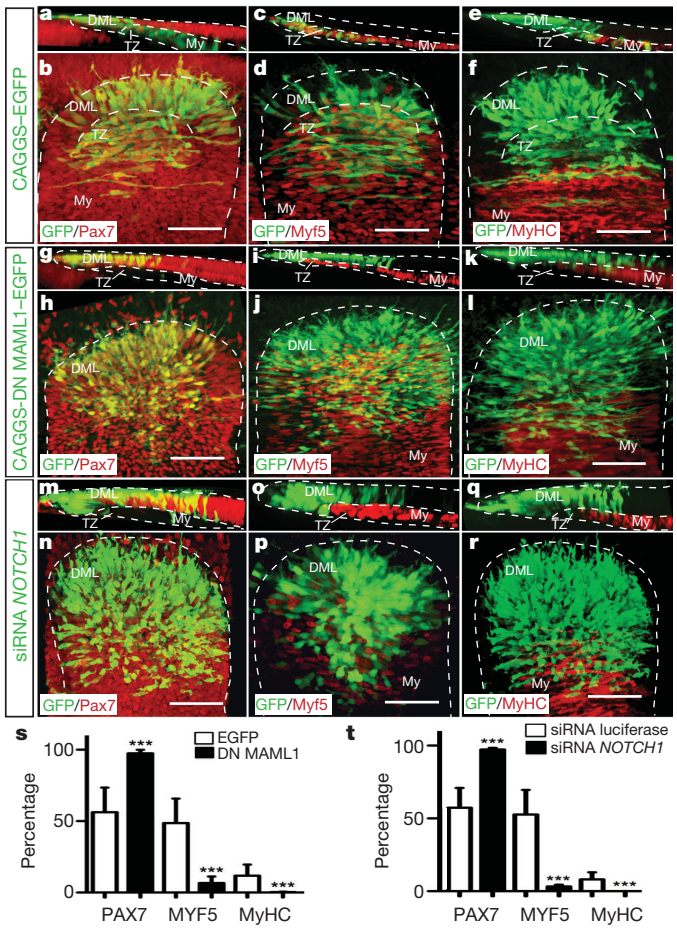


Figure 2 | NOTCH signalling is necessary for myogenesis. **a**–**r**, Confocal stacks of somites in dorsal (**b**, **d**, **f**, **h**, **j**, **l**, **n**, **p**, **r**) and transverse (**a**, **c**, **e**, **g**, **i**, **k**, **m**, **o**, **q**) view, 24 h after electroporation of (in green) CAGGS–EGFP as controls (**a**–**f**), DN MAML1 (**g**–**l**) and siRNA against chick *NOTCH1* (**m**–**r**), and stained (in red) for PAX7 (**a**, **b**, **g**, **h**, **m**, **n**), MYF5 (**c**, **d**, **i**, **j**, **o**, **p**) and MyHC (**e**, **f**, **k**, **l**, **q**, **r**). **s**, Quantification of **a**–**l**. **t**, Quantification of **m**–**r**, siRNA against luciferase as controls are not shown. Error bars show s.d. ***$P < 0.0001$. Scale bars, 50 μm.

We induced a gain of NOTCH function by electroporating NICD in newly formed somites. One day later, most electroporated cells had translocated in the transition zone (Fig. 3a–f); however, most (83%) expressed the dermomyotomal marker PAX7 (Fig. 3a, b, g), and only a few (3%) were MYF5 positive (Fig. 3c, d, g). Although few electroporated cells entered the myotome region, they did not elongate and never (0%) initiated MyHC expression (Fig. 3e–g). This result is coherent with studies that showed that NOTCH signalling inhibits muscle differentiation in various contexts[19–21]. However, characterizing the electroporated cells 6 h after electroporation of NICD, we observed a robust increase in the proportion of electroporated cells expressing MYF5 (80%; controls, 17%; Fig. 3h–k and Supplementary Fig. 5n). Strong MYF5 activation was maintained 12 h after electroporation (89%; controls, 20%; Fig. 3l–o and Supplementary Fig. 5n). After 6 h, all MYF5-positive electroporated cells were positioned in the epithelial DML (Fig. 3j), at 12 h, most electroporated cells had translocated in the transition zone (Fig. 3n). The same observations were made with MYOD (Supplementary Fig. 5a–m). Altogether, this indicates that the first steps of myogenesis (the activation of MYF5 and MYOD) are promoted by a short activation of NOTCH signalling. However, a sustained activation of NOTCH reverses the myogenic program, resulting in a downregulation of MYF5 and MYOD expression and a return to a PAX7-positive state.

To prove this, we used a doxycyclin-inducible system to drive NICD expression. In the continuous presence of doxycyclin, NICD expression was maintained in electroporated cells and, consistent with our previous observation (Fig. 3a–f), most of them translocated in the transition zone but did not maintain MYF5 expression (6%, Fig. 3r, s, v; controls, 42% MYF5-positive, Fig. 3p, q, v). When doxycyclin was removed, NICD was strongly expressed 6 h later, but was almost undetectable after overnight incubation (Supplementary Fig. 6c, d, f). Remarkably, after this transient activation of NOTCH signalling, most electroporated cells had translocated in the transition zone and the myotome and nearly all (97%) expressed MYF5 (Fig. 3t–v). In addition, electroporated cells that were positioned in the myotome had elongated into myocytes, indicating that they initiated terminal differentiation. The lack of electroporated cells in the DML (arrowheads in Fig. 3t) indicates a depletion of the DML progenitor cell population and suggests that the pulse of NOTCH signalling massively disrupted the balance between maintenance and differentiation of this cell population. This shows that NOTCH signalling displays a complex behaviour on myogenesis, acting as a potent stimulator of the myogenic program for DML cells, but only during a limited time window.

In search for a signal controlling the mosaic activation of NOTCH we observed in the DML, we noted that neural crest cells that migrate in close proximity to the DML express the NOTCH ligand Delta1 (DLL1) in a salt-and-pepper pattern (Fig. 4a and Supplementary Fig. 8a, b). A provocative hypothesis was thus that migrating DLL1-expressing neural crest cells may activate NOTCH signalling in selected progenitors within the DML. We eliminated the neural crest cell population by electroporating into the neural tube a diphtheria toxin fragment A complementary DNA (DTA) under the control of a neural-crest-specific promoter (Supplementary Fig. 7a–f). This led to a considerable decrease in the expression of MYF5 on the electroporated side (Fig. 4b, arrowheads, and Supplementary Fig. 8c; $n = 13/15$). The inhibition of non-canonical, planar cell polarity (PCP) WNT signalling in *Xenopus* affects neural crest migration[22] without affecting its induction. In the dorsal neural tube, we electroporated a mutant form of the WNT intracellular effector Dishevelled that specifically inhibits the WNT/PCP pathway[23–25]. This led to a considerable reduction in MYF5 expression compared to the control side (Fig. 4c and Supplementary

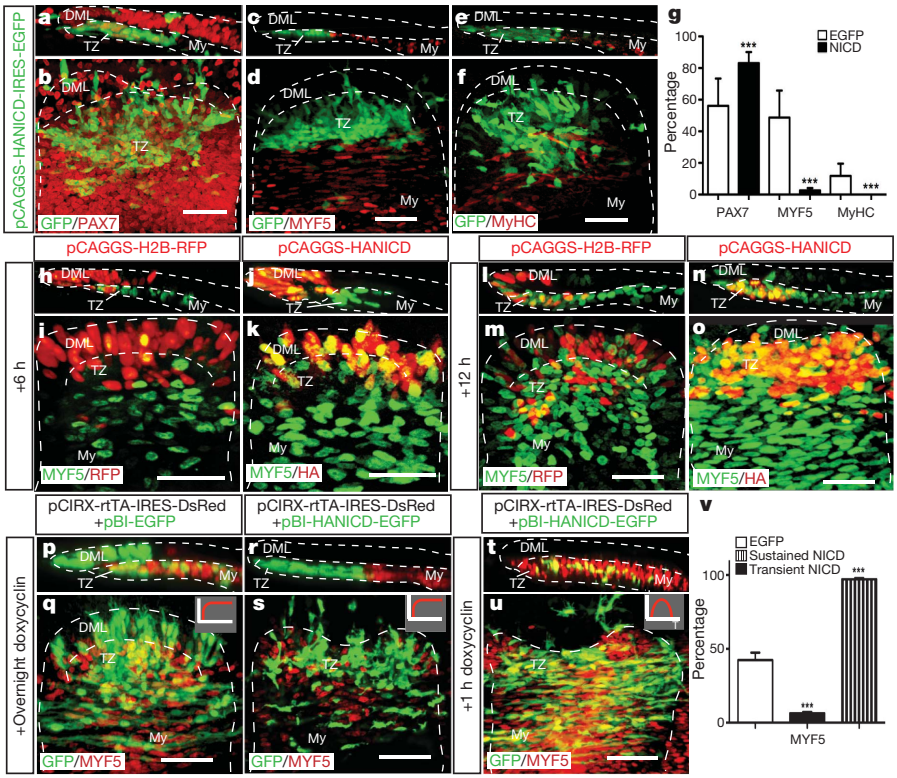


Figure 3 | Myogenesis requires the transient activation of NOTCH. **a**–**f**, Confocal stacks of somites in dorsal (**b**, **d**, **f**) and transverse (**a**, **c**, **e**) view 24 h after electroporation of NICD (in green), and stained (in red) for PAX7 (**a**, **b**), MYF5 (**c**, **d**) and MyHC (**e**, **f**). **g**, Quantification of **a**–**f**. Error bars show s.d. ***$P < 0.0001$. **h**–**o**, Time-course analysis of MYF5 expression (in green) in dorsal (**i**, **k**, **m**, **o**) and transverse (**h**, **j**, **l**, **n**) view, 6 h (**h**–**k**) and 12 h (**l**–**o**) after electroporation of CAGGS–H2B–RFP as control (**h**, **i**, **l**, **m**) or HA–NICD (**j**, **k**, **n**, **o**). In red, staining for RFP (**h**, **i**, **l**, **m**) or HA (**j**, **k**, **n**, **o**). **p**–**u**, Confocal stacks of somites in dorsal (**q**, **s**, **u**) and transverse (**p**, **r**, **t**) view electroporated with a doxycyclin-expression-inducible system. Embryos were electroporated with an empty vector as controls (**p**, **q**) or HA–NICD (**r**, **u**), treated for 1 h (**t**, **u**) or overnight (**p**, **s**) with doxycyclin and stained for GFP (in green) and MYF5 (in red). **v**, Quantification of **p**–**u**. Scale bars, 50 μm.

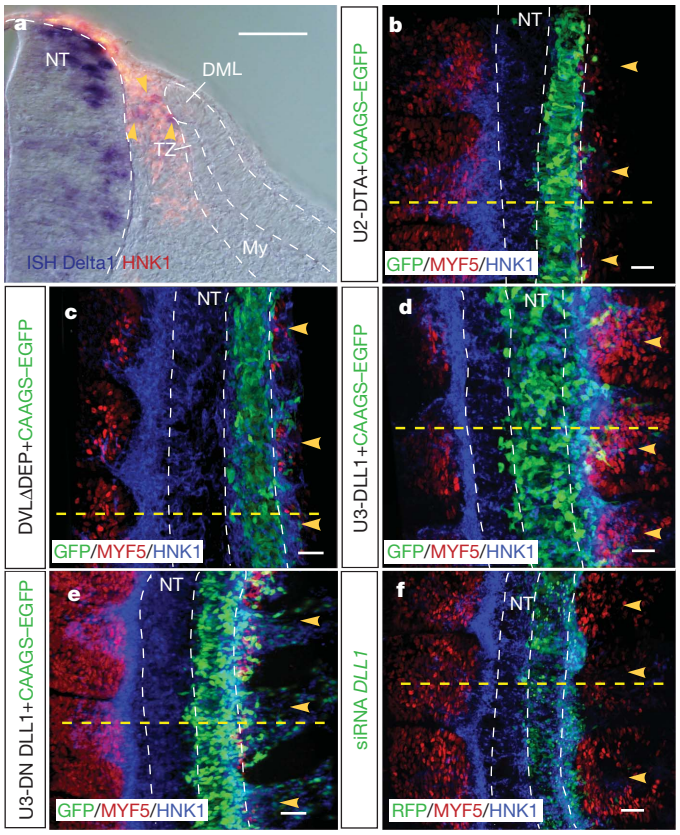


Fig. 8d; $n = 10/13$). We then electroporated DLL1 under the control of the neural-crest-specific promoter in the neural tube and verified that this resulted in the overexpression of DLL1 protein in neural crest cells (Supplementary Fig. 9a–c). We observed a significant increase of MYF5 expression (Fig. 4d and Supplementary Fig. 8e; $n = 9/13$). Loss of DLL1 function was achieved by electroporating a dominant-negative form of DLL1 (DN DLL1)[26] and with siRNAs against chick *DLL1*. DN DLL1 protein was expressed in neural crest cells (Supplementary Fig. 9k–m) and the siRNA construct efficiently reduced the endogenous DLL1 mRNA (Supplementary Fig. 9n–p) and protein (Supplementary Fig. 9q–s) levels. Both the DN DLL1 ($n = 17/17$) and the *DLL1* siRNA ($n = 10/11$) resulted in a significant reduction in MYF5 staining (Fig. 4e, f and Supplementary Fig. 8f, g). Overexpression of DLL1 in neural crest resulted in a robust activation of chick *HES1* mRNA expression (Supplementary Fig. 9d–f) and of the NOTCH reporter activity in somites (67%; Supplementary Fig. 9i, j), whereas electroporation of DTA or DN DLL1 led to a near loss of NOTCH reporter activity (1.6% and 1.8%, respectively, Supplementary Fig. 9g, h, j; controls: 11%, Fig. 1c, d; Supplementary Fig. 2l), strongly supporting the hypothesis that NOTCH ligands presented

Figure 4 | Neural crest regulates myogenesis in somites through NOTCH signalling. **a**, Mosaic expression of chick *DLL1* (in blue, yellow arrowheads) within the HNK1-positive (in red) neural crest population. **b**–**f**, Confocal stacks of neural tube, neural crest and somites in dorsal view, 24 h after electroporation of one half of the neural tube with U2-DTA (**b**), CAGGS-DVLΔDEP (**c**), U3-DLL1 (**d**), U2-DN DLL1 (**e**) (see Methods for details) and siRNA against chick *DLL1* (**f**). In green (**f**) native RFP; in green (**b**–**e**) GFP immunostaining; in red, MYF5 and in blue, HNK1. Dotted lines in **b**–**f** indicate the level of transverse sections shown in Supplementary Fig. 8c–g. Scale bars, 50 μm.

by neural crest cells modulate NOTCH signalling in somites. When DTA, DLL1 or DN DLL1 was electroporated into neural crest, MYOD expression was affected similarly to MYF5 (Supplementary Fig. 10), indicating that the two major molecular players of the early myogenic program are similarly regulated by DLL1 from neural crest. The proliferation of progenitors within the DML was not significantly changed in these experiments (Supplementary Fig. 11a–l), further supporting the hypothesis (Fig. 3t, u) that NOTCH signalling regulates the progressive differentiation of the muscle progenitor pool within the DML. Because neural crest emigrates from the neural tube during a limited time period (about 24 h from when migration initiates), the neural-crest-mediated regulation of muscle growth is limited to the initial phases of myotome formation. However, this may have long-term consequences on muscle growth, as we observed significant changes in myotome growth 48 h after electroporation of DTA, DLL1 or DN DLL1 into the neural crest, that is, 24 h after crest migration has ceased (Supplementary Fig. 12a–s). As controls, we verified that the neural crest manipulations did not affect the expression of the known modulators of myotome formation in the dorsal neural tube, that is, *WNT1*, *WNT3A* and *BMP4* (Supplementary Fig. 13a–l). It is unclear whether the same regulatory mechanisms are used in mouse. Hypomorph Dll1 mouse mutants displayed an enlarged primary myotome[20]. However, as Dll1 is expressed in both paraxial mesoderm and neural crest in early mouse embryo, the source of Notch signalling that engenders this phenotype remains to be defined. To examine this question, the inhibition of Dll1 activity in specific cell types of the mouse embryo will be required.

Our model suggests an additional role of the NOTCH pathway during myogenesis whereby, within a population of DML cells all exposed to uniform gradients of myogenic activating factors, only those cells that transiently activate the NOTCH pathway undergo myogenesis. Transient NOTCH signalling is triggered by the NOTCH ligand DLL1 carried and presented by migrating neural crest cells in a 'kiss and run' mode of signalling transduction (Supplementary Movie 2). This links the timing of myotome formation to that of neural crest migration, providing a mechanistic link for the concurrence of these two events (Supplementary Fig. 14a–g). The ability of migrating cells to influence cell fate in neighbouring tissues may reveal a general principle for generating pulses of signal activation that result in the differentiation of a defined subset of cells within a stem or progenitor pool.

METHODS SUMMARY

Electroporation, vectors, time-lapse experiments and confocal analyses. Further details can be found in Methods. The somite electroporation technique has been described elsewhere[6,27]. Time-lapse experiments were performed essentially as described[27] on transverse slices of embryos.

Quantifications and statistical analyses. On average, more than 2,300 cells were counted per point to compute the corresponding quantifications shown in Figs 1–3 and Supplementary Figs 2–6. Statistical analyses were performed using the GraphPad Prism software.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 6 September 2010; accepted 24 February 2011.
Published online 15 May 2011.

1. Shimojo, H., Ohtsuka, T. & Kageyama, R. Oscillations in Notch signaling regulate maintenance of neural progenitors. *Neuron* **58,** 52–64 (2008).
2. Joubin, K. & Stern, C. D. Molecular interactions continuously define the organizer during the cell movements of gastrulation. *Cell* **98,** 559–571 (1999).
3. Palmeirim, I., Henrique, D., Ish-Horowicz, D. & Pourquie, O. Avian hairy gene expression identifies a molecular clock linked to vertebrate segmentation and somitogenesis. *Cell* **91,** 639–648 (1997).
4. Denetclaw, W. F. Jr, Berdougo, E., Venters, S. J. & Ordahl, C. P. Morphogenetic cell movements in the middle region of the dermomyotome dorsomedial lip associated with patterning and growth of the primary epaxial myotome. *Development* **128,** 1745–1755 (2001).
5. Venters, S. J. & Ordahl, C. P. Persistent myogenic capacity of the dermomyotome dorsomedial lip and restriction of myogenic competence. *Development* **129,** 3873–3885 (2002).
6. Gros, J., Scaal, M. & Marcelle, C. A two-step mechanism for myotome formation in chick. *Dev. Cell* **6,** 875–882 (2004).
7. Kahane, N., Cinnamon, Y. & Kalcheim, C. The cellular mechanism by which the dermomyotome contributes to the second wave of myotome development. *Development* **125,** 4259–4271 (1998).
8. Kahane, N., Cinnamon, Y. & Kalcheim, C. The roles of cell migration and myofiber intercalation in patterning formation of the postmitotic myotome. *Development* **129,** 2675–2687 (2002).
9. Cinnamon, Y., Kahane, N. & Kalcheim, C. Characterization of the early development of specific hypaxial muscles from the ventrolateral myotome. *Development* **126,** 4305–4315 (1999).
10. Kahane, N., Cinnamon, Y., Bachelet, I. & Kalcheim, C. The third wave of myotome colonization by mitotically competent progenitors: regulating the balance between differentiation and proliferation during muscle development. *Development* **128,** 2187–2198 (2001).
11. Ben-Yair, R. & Kalcheim, C. Lineage analysis of the avian dermomyotome sheet reveals the existence of single cells with both dermal and muscle progenitor fates. *Development* **132,** 689–701 (2005).
12. Gros, J., Manceau, M., Thome, V. & Marcelle, C. A common somitic origin for embryonic muscle progenitors and satellite cells. *Nature* **435,** 954–958 (2005) CrossRef.
13. Relaix, F., Rocancourt, D., Mansouri, A. & Buckingham, M. A Pax3/Pax7-dependent population of skeletal muscle progenitor cells. *Nature* **435,** 948–953 (2005).
14. Kassar-Duchossoy, L. *et al.* Pax3/Pax7 mark a novel population of primitive myogenic cells during development. *Genes Dev.* **19,** 1426–1431 (2005).
15. Hirsinger, E. *et al.* Notch signalling acts in postmitotic avian myogenic cells to control MyoD activation. *Development* **128,** 107–116 (2001).
16. Fryer, C. J., Lamar, E., Turbachova, I., Kintner, C. & Jones, K. A. Mastermind mediates chromatin-specific transcription and turnover of the Notch enhancer complex. *Genes Dev.* **16,** 1397–1411 (2002).
17. Weng, A. P. *et al.* Growth suppression of pre-T acute lymphoblastic leukemia cells by inhibition of notch signaling. *Mol. Cell. Biol.* **23,** 655–664 (2003).
18. Das, R. M. *et al.* A robust system for RNA interference in the chicken using a modified microRNA operon. *Dev. Biol.* **294,** 554–563 (2006).
19. Vasyutina, E., Lenhard, D. C. & Birchmeier, C. Notch function in myogenesis. *Cell Cycle* **6,** 1450–1453 (2007).
20. Schuster-Gossler, K., Cordes, R. & Gossler, A. Premature myogenic differentiation and depletion of progenitor cells cause severe muscle hypotrophy in *Delta1* mutants. *Proc. Natl Acad. Sci. USA* **104,** 537–542 (2007).
21. Vasyutina, E. *et al. RBP-J* (*Rbpsuh*) is essential to maintain muscle progenitor cells and to generate satellite cells. *Proc. Natl Acad. Sci. USA* **104,** 4443–4448 (2007).
22. De Calisto, J., Araya, C., Marchant, L., Riaz, C. F. & Mayor, R. Essential role of non-canonical Wnt signalling in neural crest migration. *Development* **132,** 2587–2597 (2005).
23. Wallingford, J. B. *et al.* Dishevelled controls cell polarity during *Xenopus* gastrulation. *Nature* **405,** 81–85 (2000).
24. Rothbächer, U. *et al.* Dishevelled phosphorylation, subcellular localization and multimerization regulate its role in early embryogenesis. *EMBO J.* **19,** 1010–1022 (2000).
25. Gros, J., Serralbo, O. & Marcelle, C. WNT11 acts as a directional cue to organize the elongation of early muscle fibres. *Nature* **457,** 589–593 (2009).
26. Henrique, D. *et al.* Expression of a *Delta* homologue in prospective neurons in the chick. *Nature* **375,** 787–790 (1995).
27. Rios, A. C., Denans, N. & Marcelle, C. Real-time observation of Wnt β-catenin signaling in the chick embryo. *Dev. Dyn.* **239,** 346–353 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank N. Rosenthal and P. Currie for critical reading of the manuscript. This study was funded by grants from the Agence Nationale pour le Recherche (ANR), and by the EU 6th Framework Programme Network of Excellence MYORES. The help of P. Weber, S. Firth, C. Johnson and I. Harper from Imaging Facilities (IBDML, Marseille and MMI, Monash University) is acknowledged.

Author Contributions A.C.R. and C.M conceived the experiments. A.C.R. predominantly performed the work with the help of O.S. D.S. designed the animation. C.M. supervised the project and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to C.M. (christophe.marcelle@monash.edu).

METHODS

Electroporation and confocal analysis. The somite electroporation technique that was used throughout this study has been described elsewhere[6,27]. Briefly, we targeted the expression of various constructs to the dorsomedial portion of newly formed interlimb somites of Hamburger–Hamilton (HH) stage 15–16 chick embryos (24–28 somite)[28]. We have previously shown that this technique allows the specific expression of cDNA constructs in the DML of the dermomyotome[6]. To target the neural crest population, we electroporated the dorsal neural tube of HH stage 13–14 chick embryos at the level of the presomitic mesoderm.

The following constructs have been previously published: HES1–d2EGFP and the HES1–EGFP[29] (provided by R. Kageyama) contain the mouse *Hes1* promoter region upstream of destabilized or stable GFP. The CAGGS-H2B–RFP (provided by S. Tajbakhsh) contains a fusion of histone 2B with RFP downstream of the CAGGS strong ubiquitous promoter (CMV/chick β-actin promoter/enhancer). The CAGGS–EGFP[30] contains the CAGGS promoter followed by the EGFP reporter gene. The pCAB-HA-NICD-IRES-GFP (provided by N. Daudet) contains an HA-tagged NICD under the control of the CAGGS promoter[31]. The doxycyclin inducible system is composed of two plasmids that are co-electroporated: first, the pCIRX-rtTA-IRES-DsRed[32] (provided by O. Pourquié) contains a Tet-On Advanced transactivator (rtTA, Clontech) downstream of the CAGGS promoter. The IRES-DsRed-Express allows the detection of electroporated cells. Second, the pBI-HANICD-EGFP is the response plasmid (Clontech) in which the HA-tagged constitutively active form of NOTCH, NICD, was cloned. The bidirectional tetracyclin-response element drives the expressions of EGFP (which serves as an internal control of the induced response, see Supplementary Fig. 2a, b) and HANICD. pCLGFP-DVLΔDEP contains a mutated form of *Xenopus* Dishevelled that lacks the DEP domain, driven by the CAGGS promoter[25]. This construct contains also EGFP driven by its own SV40 promoter. The siRNA directed against chick *NOTCH1* has been described elsewhere[18]

We made new constructs for this study: to construct the HES1 nVENUS-PEST, a destabilized nuclear Venus GFP variant[33] was inserted downstream of the mouse *Hes1* promoter region[29]. The CAGGS-DN MAML1–EGFP contains a truncated, dominant-negative form of the human Mastermind (DN MAML1), fused with EGFP[17] downstream of the CAGGS promoter. The pCAB-HA-NICD was constructed by removing the EGFP reporter from pCAB-HA-NICD-IRES-GFP. The U2- and U3-EGFP were made by inserting the U2 and U3 evolutionary conserved *Sox10* enhancer sequences[34] in the TK-EGFP[35] plasmid, that contains the thymidine kinase minimal promoter upstream of the EGFP. The diphtheria toxin gene[36], the chick *DLL1* or a dominant-negative form of this gene[37] were inserted in the U2 or the U3-TK-EGFP in place of the EGFP to obtain the U2-DTA, the U2-DN DLL1 and the U3-DLL1 electroporation vectors. To detect electroporated cells, those plasmids were electroporated with a pCAGGS-EGFP. We have constructed two RNA interference plasmids as described previously[18] that each express two siRNAs directed against chick *DLL1*. Sequences TCACAGCGATA ACTCCGATAAA and TGCAGGAGTTTGTCAACAAGAA were inserted in siRNA chick *DLL1* A, whereas sequences GATTCAGTATATTCCACTTCAA and CCGGCACCTTCTCGCTCATCAT were inserted in siRNA chick *DLL1* B. Electroporation of plasmids A, B, or A together with B efficiently decreased the endogenous expression of chick DLL1 mRNA and protein, whereas the electroporation of siRNA directed against luciferase had no effect on chick DLL1 expression. An RFP reporter gene is inserted in the same constructs to detect electroporated cells.

Antibody stainings and BrdU labelling. For BrdU labelling, embryos were incubated for 30 min with 50 µl of a 1 mg ml^{-1} BrdU (Sigma) solution. Whole-mount antibody stainings were performed as described[25]. The following antibodies were used: rabbit polyclonals directed against chick myogenic regulatory factors MYF5 and MYOD[38]; chick DLL1[39]; and anti-RFP (Abcam); chicken polyclonals against EGFP (Abcam); rat polyclonals against the HA tag and anti-BrdU (Abcam). We also used monoclonals against the dermomyotome and dorsal neural tube marker PAX7 and against terminal myogenic differentiation marker MyHC (MF20) (obtained from the Developmental Studies Hybridoma Bank); and the neural-crest-specific monoclonal antibody HNK1 (provided by A. Eichmann).

***In situ* hybridization.** The following probes were used: chick HES1/cHairy2 (ref. 40) and chick[37] DLL1 and chick LFNG (provided by O. Pourquié), and 400 bp cDNA clones coding for fragments of chick *WNT1*, *WNT3A* and a 1 kb chick *BMP4* probe[41].

Doxycyclin-mediated induction of NOTCH signalling. Eight hours after electroporation of pCIRX-rtTA-IRES-DsRed and pBI-HANICD-EGFP, doxycyclin (300 µl of a 0.1 µg ml^{-1} solution) was added onto the embryos, and it was either washed off after one hour with PBS for transitory upregulation of NICD, or left overnight, for permanent expression of this molecule. We verified that the response plasmid is completely silent before doxycyclin addition (that is, no EGFP expression, Supplementary Fig. 6a) while it is strongly and rapidly activated 6 h after doxycyclin addition (Supplementary Fig. 6b)

Time-lapse experiments and confocal analyses. Time-lapse experiments were performed essentially as described[27] on transverse slices (250 µm) of embryos. Embryo slices were filmed for 11 h at 37 °C with a confocal inverted Leica SP5 microscope equipped with a resonant scanner, at the rate of one image stack per ten minutes. Confocal images were acquired transversally over a thickness of 100 µm; Supplementary Movie 1 corresponds to a fraction (10 µm thick) of the acquired images. Dorsal views shown in Figs 1–4 are projections of stacks of confocal images. Confocal stacks of images were visualized and analysed with the Imaris software suite. Cell countings were performed using the Improvision Volocity software suite.

Quantifications and statistical analyses. Electroporation results in the transfection of a portion of the targeted cell population, which is variable from embryo to embryo. To precisely evaluate the phenotypes obtained after electroporation of cell-autonomously acting cDNA constructs, the number of positive cells was compared to the total number of electroporated cells, recognized by an internal fluorescent reporter construct. On average, more than 2,300 cells were counted per point and the corresponding quantifications are shown in Figs 1–3 and Supplementary Figs 2–6. This mode of quantification could not be applied when constructs were electroporated in one tissue while the effects were evaluated in another, such as in experiments shown in Fig. 4 and Supplementary Figs 8–11. In this case, we report the number of embryos in which we observed a phenotype similar to the one that is illustrated in the figures, over the total number of electroporated embryos. Statistical analyses were performed using the GraphPad Prism software. Mann–Whitney non-parametric two-tail testing was applied to populations to determine the P values indicated in the figures. In each graph, columns correspond to the mean and error bars correspond to the standard deviation. ***P value < 0.0001.

28. Hamburger, V. & Hamilton, H. L. A series of normal stages in the development of the chick embryo. *Dev. Dyn.* **195,** 231–272 (1992).
29. Ohtsuka, T. *et al.* Visualization of embryonic neural stem cells using Hes promoters in transgenic mice. *Mol. Cell. Neurosci.* **31,** 109–122 (2006).
30. Tobiume, M. *et al.* Inefficient enhancement of viral infectivity and CD4 downregulation by human immunodeficiency virus type 1 Nef from Japanese long-term nonprogressors. *J. Virol.* **76,** 5959–5965 (2002).
31. Daudet, N. & Lewis, J. Two contrasting roles for Notch activity in chick inner ear development: specification of prosensory patches and lateral inhibition of hair-cell differentiation. *Development* **132,** 541–551 (2005).
32. Iimura, T. & Pourquie, O. Collinear activation of *Hoxb* genes during gastrulation is linked to mesoderm cell ingression. *Nature* **442,** 568–571 (2006).
33. Nagoshi, E. *et al.* Circadian gene expression in individual fibroblasts: cell-autonomous and self-sustained oscillators pass time to daughter cells. *Cell* **119,** 693–705 (2004).
34. Werner, T., Hammer, A., Wahlbuhl, M., Bosl, M. R. & Wegner, M. Multiple conserved regulatory elements with overlapping functions determine Sox10 expression in mouse embryogenesis. *Nucleic Acids Res.* **35,** 6526–6538 (2007).
35. Uchikawa, M., Ishida, Y., Takemoto, T., Kamachi, Y. & Kondoh, H. Functional analysis of chicken *Sox2* enhancers highlights an array of diverse regulatory elements that are conserved in mammals. *Dev. Cell* **4,** 509–519 (2003).
36. Maxwell, I. H., Maxwell, F. & Glode, L. M. Regulated expression of a diphtheria toxin A-chain gene transfected into human cells: possible strategy for inducing cancer cell suicide. *Cancer Res.* **46,** 4660–4664 (1986).
37. Henrique, D. *et al.* Maintenance of neuroepithelial progenitor cells by Delta–Notch signalling in the embryonic chick retina. *Curr. Biol.* **7,** 661–670 (1997).
38. Manceau, M. *et al.* Myostatin promotes the terminal differentiation of embryonic muscle progenitors. *Genes Dev.* **22,** 668–681 (2008).
39. Henrique, D. *et al.* cash4, a novel achaete-scute homolog induced by Hensen's node during generation of the posterior nervous system. *Genes Dev.* **11,** 603–615 (1997).
40. Jouve, C. *et al.* Notch signalling is required for cyclic expression of the hairy-like gene *HES1* in the presomitic mesoderm. *Development* **127,** 1421–1429 (2000).
41. Marcelle, C., Stark, M. R. & Bronner-Fraser, M. Coordinate actions of BMPs, Wnts, Shh and noggin mediate patterning of the dorsal somite. *Development* **124,** 3955–3963 (1997).

LETTER

doi:10.1038/nature09956

Structure of the spliceosomal U4 snRNP core domain and its implication for snRNP biogenesis

Adelaine K. W. Leung[1]†, Kiyoshi Nagai[1] & Jade Li[1]

The spliceosome is a dynamic macromolecular machine that assembles on pre-messenger RNA substrates and catalyses the excision of non-coding intervening sequences (introns)[1–3]. Four of the five major components of the spliceosome, U1, U2, U4 and U5 small nuclear ribonucleoproteins (snRNPs), contain seven Sm proteins (SmB/B′, SmD1, SmD2, SmD3, SmE, SmF and SmG) in common[4,5]. Following export of the U1, U2, U4 and U5 snRNAs to the cytoplasm[6,7], the seven Sm proteins, chaperoned by the survival of motor neurons (SMN) complex, assemble around a single-stranded, U-rich sequence called the Sm site in each small nuclear RNA (snRNA), to form the core domain of the respective snRNP particle[8,9]. Core domain formation is a prerequisite for re-import into the nucleus[10], where these snRNPs mature via addition of their particle-specific proteins. Here we present a crystal structure of the U4 snRNP core domain at 3.6 Å resolution, detailing how the Sm site heptad (AUUUUUG) binds inside the central hole of the heptameric ring of Sm proteins, interacting one-to-one with SmE–SmG–SmD3–SmB–SmD1–SmD2–SmF. An irregular backbone conformation of the Sm site sequence combined with the asymmetric structure of the heteromeric protein ring allows each base to interact in a distinct manner with four key residues at equivalent positions in the L3 and L5 loops of the Sm fold. A comparison of this structure with the U1 snRNP at 5.5 Å resolution[11,12] reveals snRNA-dependent structural changes outside the Sm fold, which may facilitate the binding of particle-specific proteins that are crucial to biogenesis of spliceosomal snRNPs.

Proteins in the Sm family are characterized by Sm1 and Sm2 motifs joined by a variable linker[13–15] (Supplementary Fig. 1). SmB/B′, SmD1 and SmD3 contain extended C termini, whereas SmD2 and SmE contain extended N termini. The Sm-fold consists of an N-terminal α-helix and a five-stranded antiparallel β-sheet containing the Sm motifs and folded upon itself[16] (Supplementary Fig. 2). The subunit interfaces in the SmD1–SmD2 and SmD3–SmB heterodimers suggest that the seven Sm proteins are assembled in a ring[16] in the snRNP core domain[17,18], and this has been confirmed by the crystal structure of the U1 snRNP at 5.5 Å resolution[11,12]. Crystal structure of *Archaeoglobus fulgidus* Lsm-1 homo-heptamer in complex with penta-uridylate showed how Lsm-1 provides U-specificity[19,20]. However, crystal structures of U1 snRNP at 5.5 or 4.4 Å resolution have left the side chain interactions between the Sm site and Sm proteins unresolved[11,12,21]. Hence, it remains unknown how the heteromeric Sm proteins combine to specifically recognize the Sm site sequence.

We solved the structure of the U4 snRNP core domain (Supplementary Table 1) assembled on a fragment of U4 snRNA (Supplementary Figs 3 and 4), which crystallized[22] with 12 copies in the asymmetric unit (Supplementary Fig. 5). The seven Sm proteins form a ring with a relatively flat face over which the N-terminal helices lie, and a tapered face carrying the L4 loops (Fig. 1a). U4 snRNA threads through the central hole lined by loops L3, L5 and L2 (Fig. 1c). The Sm site sequence is bound around the inner wall near the rim on the flat face. Its phosphates are exposed in the hole (Fig. 1b), revealing an irregular backbone conformation (Fig. 2), and its bases project into the Sm proteins and are bound by their loops L3 and L5; they vary in orientation from nearly parallel to nearly perpendicular to the ring plane (Fig. 2 and Supplementary Fig. 6). The 5′ flanking adenine (A118) is outside the hole, and the 5′-stem (U4 SL-II) makes little contact with the Sm proteins (Fig. 1a). The 3′ flanking nucleotides (A126–C127) traverse the hole with phosphates contacting SmD1 and SmB, and bases contacting L2 of SmD2 and SmF and L5 of SmF. As the 3′-stem (U4 SL-III) emerges on the tapered face with the helical axis roughly 60° to the plane of the ring, the first base pair (C127:G144) comes into contact with Trp 25 of SmF (Trp 25F) in loop L2. The 3′ unpaired nucleotide (G145) is wedged between SmE and SmF. Along the 3′-stem the phosphates on both strands interact with basic residues from L2 and L4 loops of all the Sm proteins except SmG, particularly those from the lysine-rich, long L4 loops of SmD2 and SmB (Supplementary Fig. 7). These interactions can support the association between Sm proteins and different snRNAs during core domain assembly[23,24].

Small shape differences over the β-sheet of different Sm proteins (Supplementary Table 2), attributable to the size differences of the conserved inward pointing residues between the corresponding Sm1 and Sm2 motifs (Supplementary Fig. 1), cause the Sm protein ring to be asymmetric. The L3 and L5 loops are held at different heights and orientations relative to the plane of the ring (Supplementary Fig. 8), allowing their key residues to contact each nucleotide uniquely. The nucleotide-binding loops L3 and L5 have the consensus sequence of Asp-hydrophilic-aromatic-Met-Asn (residues L3.1–L3.5) and Arg-Gly-(acidic/Asn) (residues L5.1–L5.3) (Supplementary Table 3). In the co-crystal structure of *A. fulgidus* Lsm-1 with penta-uridylate, U specificity is achieved by sandwiching the uridine base between the side chains of His 37(L3.3) and Arg 63(L5.1), and hydrogen-bonding of its N3 and O4 atoms respectively with Oδ1 and Nδ2 of the invariant Asn 39(L3.5)[19]. In the U4 core domain (Fig. 3), however, the base stacking with the aromatic residue (L3.3) present in five of the Sm proteins, and the interaction with the Arg/Lys (L5.1), vary along the heptad and from the Lsm-1 example (Fig. 3 and Supplementary Fig. 9), due to the irregular RNA conformation around the central hole (Fig. 2 and Supplementary Fig. 6) and the sequence variations in L3 and L5 (Supplementary Table 3). We have inferred hydrogen-bonding interactions around the Sm site from the residue positions, corroborated by unambiguous orientation of base planes and aromatic side chains in the electron density, and the non-crystallographic symmetry (ncs) agreement in the conformation of the conserved contact residues (Fig. 3).

At the first position of the heptad an adenine is required, because replacement by G abolished core domain assembly with a U4 oligonucleotide, and replacement by U destabilized the assembly[23]. The A119 base is stacked with Tyr 53E(L3.3) and hydrogen-bonded at N3 to Lys 80E(L5.1), at N6 to Asp 51E(L3.1) and at N1 to Asn 55E(L3.5) (Fig. 3a), demonstrating A-specificity of its binding pocket. A salt bridge between its phosphate and Arg 61D2, which is hydrogen-bonded to

[1]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK. †Present address: Department of Neurobiology, Harvard Medical School, 220 Longwood Avenue, Boston, Massachusetts 02115, USA.

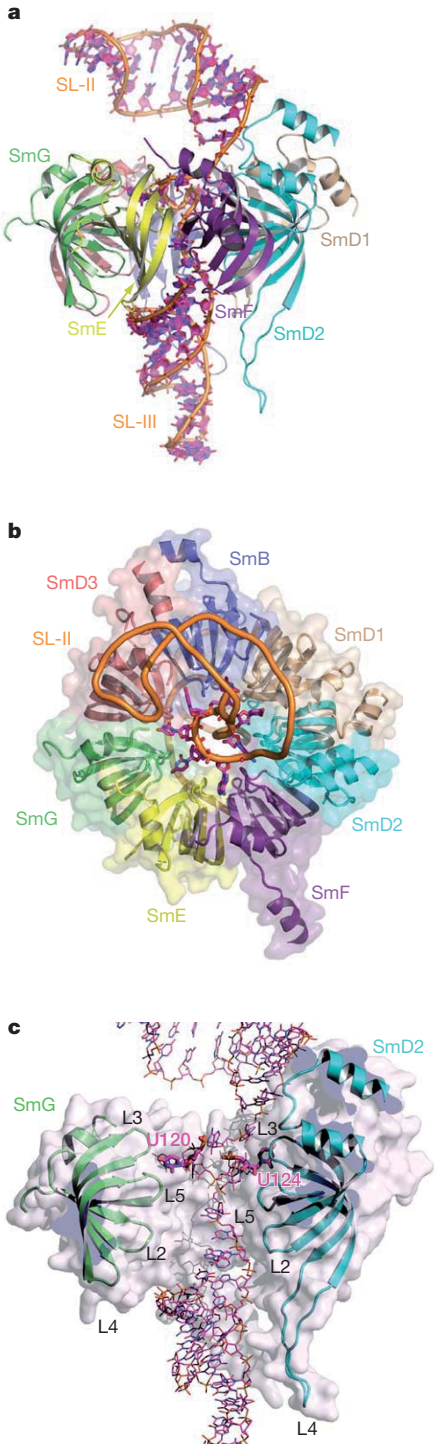


Figure 1 | Overall structure of the U4 snRNP core domain. **a**, Side view of the core domain showing the ring with its flat face up and tapered face down. **b**, View from the flat face of the ring. The N- and C-terminal extensions of the Sm fold interact between SmD3 and SmB, and between SmD1 and SmD2. **c**, The heptameric ring is cleaved along a plane (dark blue patches) through SmG and SmD2, leaving the five subunits SmG–SmD3–SmB–SmD1–SmD2 that bind the penta-uridylate to form the protein envelope in the background. Loops L3, L5 and L2 of the Sm fold line the walls of the funnel shaped hole, whereas L4 is exposed on the tapered face. The bases of the Sm site nucleotides, such as U120 and U124, are bound between L3 and L5 near the rim on the flat face.

Asp 37F (Fig. 3g), stabilizes A119 binding to the ring. In U1 snRNP and the reconstituted U4 core domain, the N7 atom of this adenine is unexpectedly nucleophilic in becoming methylated by dimethylsulphate. It indicates a perturbation of the π-electron distribution over

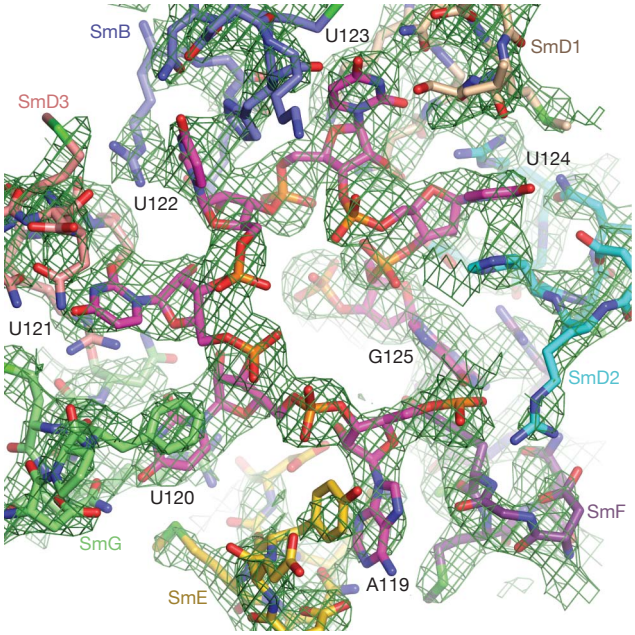


Figure 2 | The Sm site RNA binds asymmetrically in the central hole of the heptamer ring. The Sm site heptad sequence is shown together with the contact residues on Sm proteins around the central hole of the core domain. Carbon atoms are colour-coded by chain: SmE (yellow), SmG (green), SmD3 (salmon), SmB (light blue), SmD1 (tan), SmD2 (cyan), SmF (purple), RNA (magenta). Nitrogen, oxygen, sulphur and phosphorus atoms are in blue, red, green and orange, respectively. A sharpened ($B = -15$ Å^2) ncs-averaged electron density map is contoured at 8σ.

its double ring system[25], which could result from multiple hydrogen bonds to the base.

The U120 base is sandwiched between the side chains of Phe 37G and Arg 63G and hydrogen-bonded at O4 with Nδ2 of Asn 39G (Fig. 3b). Similarly, the U122 base is stacked with His 37B and hydrogen-bonded at O4 to Asn 39B (Fig. 3d), except that a different base orientation relative to L3 requires His 37B to adopt a rotamer different from that of His 62D2 (Supplementary Fig. 9). These interactions are similar to the Lsm-1 complex[19] and account for the cross-linking of the first and third U to L3 residues of SmG and SmB, respectively[26]. SmD3 and SmD1 lack the aromatic residue at L3.3. SmD3 displays U specificity in a novel mode distinct from the LSm-1 complex (ref. 19): U121 is hydrogen-bonded on O4 to both Asn 38D3 and Asn 40D3, on N3 to Asn 40D3 and on O2 to the peptide amide of Ser 66D3, besides being stacked with Arg 64D3 (Fig. 3c). U123 forms no stacking interaction (Fig. 3e). Its base is within hydrogen-bonding distance from the side chain of the invariant Asn 37D1, which is positioned to form the conserved buttressing hydrogen bonds with Asp 33D1[16,19,20]. This configuration is consistent with U123 adopting an enol tautomer that would present O2(H) and N3 as the hydrogen-bond donor and acceptor, respectively, to Oδ1 and Nδ2 of Asn 37D1. Consequently, U123 is accommodated without U-specific base contacts, which explains the lack of preference for U at this position (Supplementary Fig. 3e). In human U1 snRNA G replaces U, and in U1 snRNA of other species all four bases are tolerated (http://rfam.sanger.ac.uk)[27]. U124 is hydrogen-bonded on O2, not O4, to the invariant Asn 64D2 (Fig. 3f), and therefore our structure cannot fully account for the preference for U at this position. His 62D2 stacks with U124 and in some ncs copies also interacts edge-to-face with A118, the 5′-flanking adenine (not shown). The G125 base is not intimately associated with SmF, as it is stacked only on one edge between Tyr 39F and Arg 65F, and is too distant for hydrogen bonding with the invariant Asn 41F (Fig. 3g). Cys 66F, which is absolutely conserved in SmF, replaces the Gly in L5 without causing a clash. The absence of G-specific base contacts explains why replacing this G with A had no effect on Sm protein

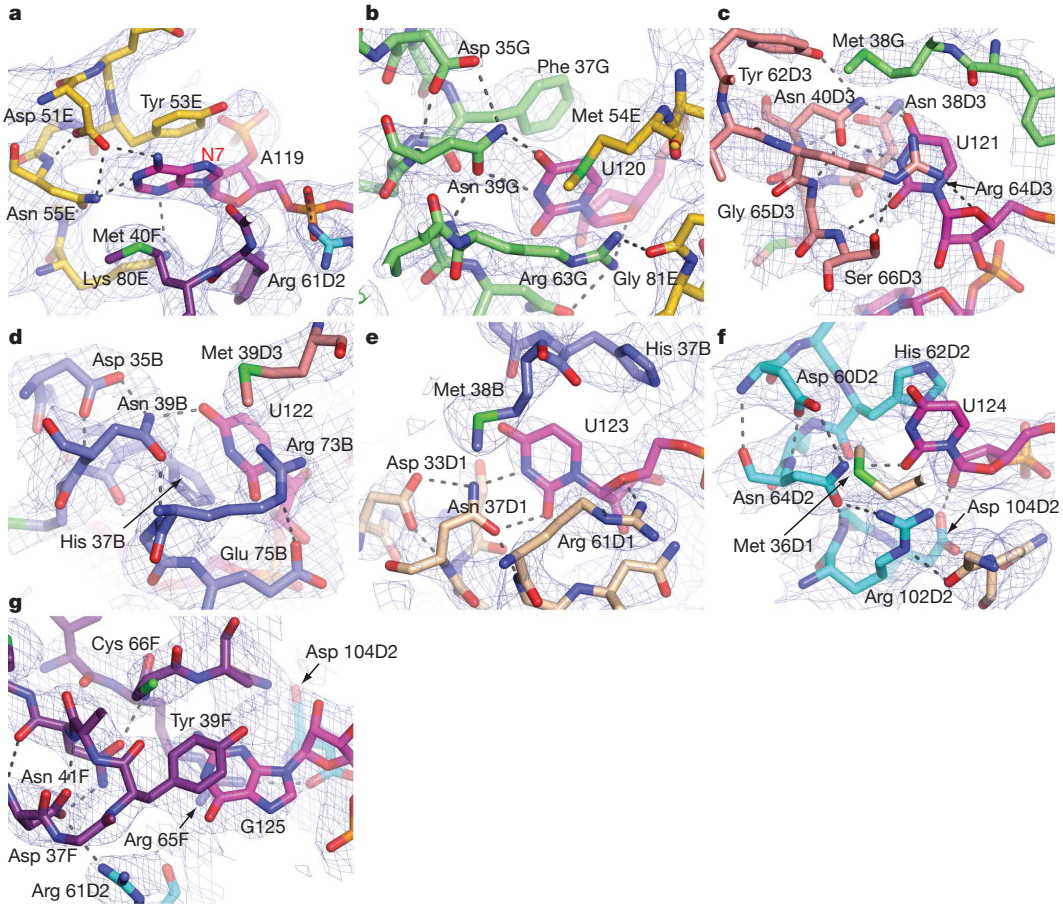


Figure 3 | Interactions between the U4 Sm site heptad nucleotides and the Sm proteins. **a**, A119. **b**, U120. **c**, U121. **d**,U122. **e**, U123. **f**, U124. **g**, G125. Shown in dashed lines are the hydrogen-bonding interactions inferred from the residue configurations. Similarities with the RNA-free heterodimers[16] and the Lsm-1 complex[19] are evident: the invariant Asn(L3.5) is buttressed by hydrogen bonds with the side chain of Asp(L3.1) and peptide amide of Gly/Cys(L5.2) in six cases (**a**, **b**, **d**–**g**) and with Glu 36D3(L3.1) via Tyr 62D3 and with the peptide amide of Gly 65D3(L5.2) in SmD3 (**c**); Met(L3.4) contributes van der Waals contacts to the base bound by a neighbouring Sm protein in six cases (**a**–**f**), and Gly(L5.2) is conserved in six cases (**a**–**f**) where a side chain would clash with the base contacting residues. The $2F_o - F_c$ map is shown sharpened with $B = -15$ Å^2 and contoured at ~1.5σ. Atom colours are as in Fig. 2.

binding[23], which is consistent with replacement of this G (Supplementary Fig. 3e) by other bases in U4 and U5 snRNAs of different species[27,28]. Thus, the last nucleotide of the Sm site heptad acts as a transition to the variable 3′-stem. In the bound heptad the phosphate groups of U120, U121 and U122 are in close proximity. Their negative charge density is likely stabilized by electrostatically held magnesium

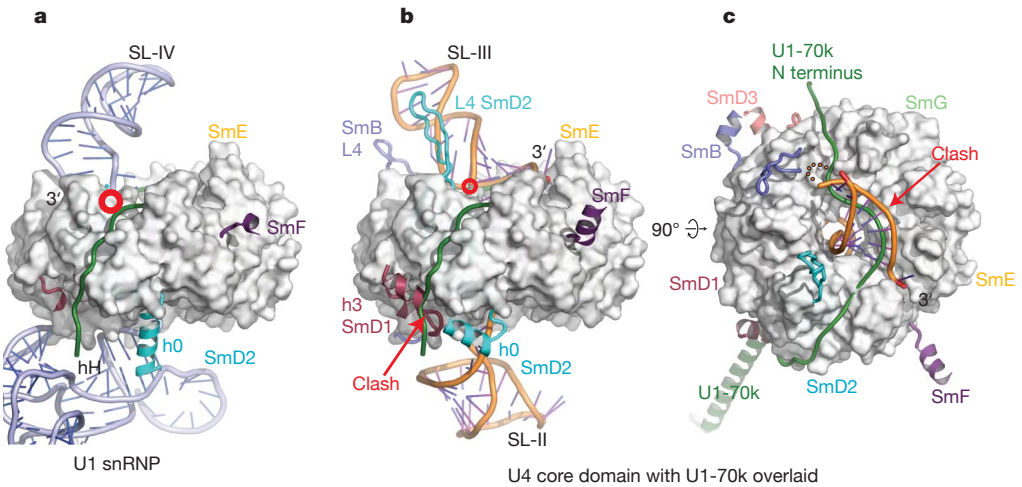


Figure 4 | snRNA-dependent structural changes of the U1 and U4 core domains. **a**, U1 snRNP[11] and **b**, U4 core domain in the same orientation; **c**, the U4 core domain in tapered-face view. The N-terminal fragment of U1-70K (green)[11,12] is overlaid onto the U4 core domain (**b** and **c**), and the Sm folds common to both structures are masked with a white envelope. The red circle indicates where U1 and U4 snRNAs emerge from the central hole. In U4, the L4 loops of SmB and SmD2 contact the backbone of the 3′-stem (**b**). In U1[11,21], SmD2 helix h0 points into the minor groove of RNA helix H (hH) (**a**), but in U4, SmD2 h0 is orientated almost orthogonal to this, with its N terminus pointing at SmD1 helix h3 (**b**). The latter is positioned to obstruct the path of U1-70K (arrow). Moreover the first seven base pairs of the 3′-stem of U4 snRNA (orange) would clash on its 3′-strand with U1-70K (arrow) (**c**).

ions of indeterminate positions[29] or by chelated cations unresolvable at our resolution.

In mammalian U4 and U5 snRNAs the Sm site heptad and the 3′-stem are linked by a single nucleotide, whereas in U1 and U2 snRNA they are linked by five nucleotides[27,28]. Structural comparison between U1 snRNP[11,12] and the U4 core domain shows that the snRNAs emerge from the central hole in similar positions (red circle in Fig. 4a and b) but with different stem orientations that cause their 3′ termini to fall on opposite sides of the hole. In U1 snRNP, the N-terminal 60 residues of U1-70K wrap around the tapered face of the core domain by skirting around the RNA stalk passing SmD3–SmG–SmE–SmF–SmD2[11,12]. Mapping the U1-70K N-terminal fragment onto the U4 core domain shows that the 3′ strand of U4 SL-III would obstruct its polypeptide path (Fig. 4c) and hence prevent its binding to the snRNP. The N-terminal 97-residue fragment of U1-70K is sufficient to bind to the U1 core domain, but fails to bind to the U5 core domain[30]. In U5 snRNA the 3′-stem is linked to the Sm site as closely as in U4 (Supplementary Fig. 3c, d) and could exclude U1-70K analogously.

The N-terminal extension of SmD2 and C-terminal extension of SmD1 are disordered in the absence of RNA[16]. In U1 snRNP, SmD2 forms an extra helix (helix 0) at the N terminus that points into the minor groove of RNA helix H (Fig. 4a). This anchors the SmD2 helix 1, whose C terminus interacts with U1-70K[11,12,21]. In the U4 core domain, the C-terminal extension of SmD1 forms helix 3, whereas the N-terminal extension of SmD2 forms helix 0 (Fig. 4b) in the ncs copies where the loop between SmD2 helices 0 and 1 interacts with the backbone of the 5′-stem (Fig. 1a). The SmD1 helix 3 interacting with the SmD2 helix 1 in its U4 snRNA-dependent orientation might also obstruct U1-70K (Fig. 4b). In U1 snRNP, the SmB helix 1 interacts with the backbone of stem II (ref. 11). These snRNA-dependent structural differences on the flat face of the core domain may, in addition to the snRNA itself, provide selectivity for the cognate particle-specific proteins and have a crucial role in snRNP biogenesis.

METHODS SUMMARY

The U4 snRNP core domain was reconstituted from the seven Sm proteins and a variant of human U4 snRNA[16,22]. Crystals in space group $P3_1$ with 12 complexes per asymmetric unit were grown by vapour diffusion and diffracted X-rays anisotropically to 3.45 Å resolution. Initial phases were determined by the multiwavelength anomalous diffraction (MAD) method using SeMet substitution within Sm sub-complexes. The structure containing 8,101 protein and RNA residues has been refined under 12-fold ncs restraints at 66.2–3.6 Å resolution to R_{free} of 32.1% with excellent geometry (Supplementary Table 1).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 2 September 2010; accepted 17 February 2011.
Published online 24 April 2011.

1. Burge, C. B., Tuschl, T. & Sharp, P. A. in *The RNA World* 2nd edn (eds Gesteland, R. R., Cech, T. R. & Atkins, J. F.) 525–560 (Cold Spring Harbor Laboratory Press, 1999).
2. Will, C. L. & Lührmann, R. in *The RNA World* 3rd edn (eds Gesteland, R. F., Cech, T. R. & Atkins, J. F.) 369–400 (Cold Spring Harbor Laboratory Press, 2006).
3. Yu, Y.-T., Scharl, E. C., Smith, C. M. & Steitz, J. A. in *The RNA World* 2nd edn (eds Gesteland, R. R., Cech, T. R. & Atkins, J. F.) 487–524 (Cold Spring Harbor Laboratory Press, 1999).
4. Hinterberger, M., Pettersson, I. & Steitz, J. A. Isolation of small nuclear ribonucleoproteins containing U1, U2, U4, U5, and U6 RNAs. *J. Biol. Chem.* **258,** 2604–2613 (1983).
5. Bringmann, P. & Lührmann, R. Purification of the individual snRNPs U1, U2, U5 and U4/U6 from HeLa cells and characterization of their protein constituents. *EMBO J.* **5,** 3509–3516 (1986).
6. Mattaj, I. W. Cap trimethylation of U snRNA is cytoplasmic and dependent on U snRNP protein binding. *Cell* **46,** 905–911 (1986).
7. Ohno, M., Segref, A., Bachi, A., Wilm, M. & Mattaj, I. W. PHAX, a mediator of U snRNA nuclear export whose activity is regulated by phosphorylation. *Cell* **101,** 187–198 (2000).
8. Meister, G., Eggert, C. & Fischer, U. SMN-mediated assembly of RNPs: a complex story. *Trends Cell Biol.* **12,** 472–478 (2002).
9. Pellizzoni, L., Yong, J. & Dreyfuss, G. Essential role for the SMN complex in the specificity of snRNP assembly. *Science* **298,** 1775–1779 (2002).
10. Fischer, U., Sumpter, V., Sekine, M., Satoh, T. & Lührmann, R. Nucleo-cytoplasmic transport of U snRNPs: definition of a nuclear location signal in the Sm core domain that binds a transport receptor independently of the m3G cap. *EMBO J.* **12,** 573–583 (1993).
11. Pomeranz Krummel, D. A., Oubridge, C., Leung, A. K. W., Li, J. & Nagai, K. Crystal structure of the human spliceosomal U1 snRNP at 5.5 Å resolution. *Nature* **458,** 475–480 (2009).
12. Oubridge, C., Pomeranz Krummel, D. A., Leung, A. K. W., Li, J. & Nagai, K. Interpreting a low resolution map of human U1 snRNP using anomalous scatterers. *Structure* **17,** 930–938 (2009).
13. Hermann, H. *et al.* snRNP Sm proteins share two evolutionarily conserved sequence motifs which are involved in Sm protein–protein interactions. *EMBO J.* **14,** 2076–2088 (1995).
14. Seraphin, B. Sm and Sm-like proteins belong to a large family: identification of proteins of the U6 as well as the U1, U2, U4 and U5 snRNPs. *EMBO J.* **14,** 2089–2098 (1995).
15. Cooper, M., Johnston, L. H. & Beggs, J. D. Identification and characterization of Uss1p (Sdb23p): a novel U6 snRNA-associated protein with significant similarity to core proteins of small nuclear ribonucleoproteins. *EMBO J.* **14,** 2066–2075 (1995).
16. Kambach, C. *et al.* Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. *Cell* **96,** 375–387 (1999).
17. Kastner, B. & Lührmann, R. Electron microscopy of U1 small nuclear ribonucleoprotein particles: shape of the particle and position of the 5′ RNA terminus. *EMBO J.* **8,** 277–286 (1989).
18. Kastner, B., Bach, M. & Lührmann, R. Electron microscopy of small nuclear ribonucleoprotein (snRNP) particles U2 and U5: evidence for a common structure-determining principle in the major U snRNP family. *Proc. Natl Acad. Sci. USA* **87,** 1710–1714 (1990).
19. Törö, I. *et al.* RNA binding in an Sm core domain: X-ray structure and functional analysis of an archaeal Sm protein complex. *EMBO J.* **20,** 2293–2303 (2001).
20. Thore, S., Mayer, C., Sauter, C., Weeks, S. & Suck, D. Crystal structures of the *Pyrococcus abyssi* Sm core and its complex with RNA. Common features of RNA binding in Archaea and Eukarya. *J. Biol. Chem.* **278,** 1239–1247 (2003).
21. Weber, G., Trowitzsch, S., Kastner, B., Lührmann, R. & Wahl, M. C. Functional organization of the Sm core in the crystal structure of human U1 snRNP. *EMBO J.* **29,** 4172–4184 (2010).
22. Leung, A. K. W. *et al.* Use of RNA tertiary interaction modules for the crystallization of the spliceosomal snRNP core domain. *J. Mol. Biol.* **402,** 154–164 (2010).
23. Raker, V. A., Hartmuth, K., Kastner, B. & Lührmann, R. Spliceosomal U snRNP core assembly: Sm proteins assemble onto an Sm site RNA nonanucleotide in a specific and thermodynamically stable manner. *Mol. Cell. Biol.* **19,** 6554–6565 (1999).
24. McConnell, T. S., Lokken, R. P. & Steitz, J. A. Assembly of the U1 snRNP involves interactions with the backbone of the terminal stem of U1 snRNA. *RNA* **9,** 193–201 (2003).
25. Hartmuth, K., Raker, V. A., Huber, J., Branlant, C. & Lührmann, R. An unusual chemical reactivity of Sm site adenosines strongly correlates with proper assembly of core U snRNP particles. *J. Mol. Biol.* **285,** 133–147 (1999).
26. Urlaub, H., Raker, V. A., Kostka, S. & Lührmann, R. Sm protein–Sm site RNA interactions within the inner ring of the spliceosomal snRNP core structure. *EMBO J.* **20,** 187–196 (2001).
27. Gardner, P. P. *et al.* Rfam: updates to the RNA families database. *Nucleic Acids Res.* **37,** D137–D140 (2009).
28. Guthrie, C. & Patterson, B. Spliceosomal snRNAs. *RNA* **9,** 193–201 (2003).
29. Draper, D. E. RNA folding: thermodynamic and molecular descriptions of the roles of ions. *Biophys. J.* **95,** 5489–5495 (2008).
30. Nelissen, R. L., Will, C. L., van Venrooij, W. J. & Lührmann, R. The association of the U1-specific 70K and C proteins with U1 snRNPs is mediated in part by common U snRNP proteins. *EMBO J.* **13,** 4113–4125 (1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the Medical Research Council of the UK and a HFSP grant. A.K.W.L. was supported by the Overseas Research Students Awards Scheme, Canada-Cambridge Commonwealth studentship, a postgraduate scholarship from NSERC and a Junior Research Fellowship from Sidney Sussex College, Cambridge University. We thank the European Synchrotron Radiation Facility and Daresbury beamline staff for their support. We thank M. Jinek, M. Kampmann and Y. Kondo for their help with crystallization. We also thank C. Kambach, J. Avis, R. Young, S. Walke and H. Teo for laying the foundation of this project, C. Oubridge and D. Pomeranz Krummel for sharing Sm proteins and providing help and advice throughout the project, and P. Zwart for advice on twinning.

Author Contributions A.K.W.L. and K.N. designed the constructs. A.K.W.L. crystallized the core domain, collected data and solved the structure in $P6_122$. J.L. identified twinning and refined the structure in $P3_1$. All three authors wrote the paper.

Author Information Atomic coordinates and structure factors for the U4 snRNP core domain have been deposited in the PDB data bank under accession numbers 2Y9A, 2Y9B, 2Y9C and 2Y9D. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to K.N. (kn@mrc-lmb.cam.ac.uk) and J.L. (jl@mrc-lmb.cam.ac.uk).

METHODS

Crystallization. Human Sm proteins were overproduced as two heterodimers, SmD1–SmD2 and SmD3–SmB, and a trimer, SmE–SmF–SmG, in *Escherichia coli* as described previously[16,22]. The Sm proteins expressed were full-length except for SmB, whose extended C terminus was truncated to residue 95. The 3′ fragment of human U4 snRNA from residues 85–145, comprising the Sm site and both flanking helices (Supplementary Fig. 4a), was produced by *in vitro* transcription[22,31]. As the U4 Sm site is almost immediately flanked by helices (Supplementary Fig. 4a), this fragment was thought likely to form a compact assembly with the Sm proteins. To promote crystal contacts, we replaced residues 97–104 of the native sequence by a stable GAAA tetraloop and residues 134–137 by a GAAA tetraloop receptor (Supplementary Fig. 4b). The human U4 core domain was reconstituted from the Sm proteins and the modified 3′ fragment of U4 snRNA, and purified on a monoQ column. Crystals were obtained by sitting drop vapour diffusion using a reservoir solution containing 4–10% (v/v) PEG 550 MME, 0.1 M KSCN or $(NH_4)_2SO_4$, 0.1 M Tris-HCl pH 8.0–8.5, 10 mM $MgCl_2$, and 2 mM cyclen (a polyamine)[32]. A full account of our crystallization effort has been published elsewhere[22].

Preparation of Se-Met derivatives. Se-Met labelled SmD1–SmD2 and SmD3–SmB heterodimers were expressed by the inhibition method[33]. Se-Met labelled SmE–SmF–SmG hetero-trimer was expressed in the methionine auxotroph B834, grown with the Se-Met core medium (Wako, 391-01541) as supplement. Incorporation of Se-Met was verified by electrospray time-of-flight mass spectrometry of the purified proteins. By replacing one or all of the heterodimers and trimer with the Se-Met-labelled subcomplexes during the reconstitution, we obtained differently labelled Se-Met derivatives of the core domain.

Data processing. Data from Se-Met derivatives were processed using MOSFLM[34], with anisotropic resolution limits when required, SCALA[35] and TRUNCATE[36]. Native data were integrated using XDS[37], converted to the CCP4 format using POINTLESS[35], and scaled using SCALA. Between randomly partitioned half-sets of native data the correlation[35] was 92% at 4.1 Å, decreasing to ~50% at 3.6 Å, and therefore data to 3.6 Å were included in the refinement (Supplementary Table 1). Before refinement, the native amplitudes were corrected for anisotropy with truncation in the weak direction to $F/\sigma F \geq 3$ (www.doe-mbi.ucla.edu/~sawaya/anisoscale/)[38].

The crystals are of space group $P3_1$ with 12 copies of the core domain in the asymmetric unit related by *222* rotational non-crystallographic symmetry (ncs) and threefold translational non-crystallographic symmetry in the *a,b*-plane (translational pseudo-symmetry). The crystals are twinned along the *a,b*- and *a*,b**-axes, but twinning was masked by the pseudo-symmetry and anisotropy, resulting in apparent $P6_122$ symmetry for data sets at low resolution. The pseudo-symmetry gives rise to self-Patterson vectors at (1/3, 2/3, 0) and (2/3, 1/3, 0), and causes intensity modulations up to 5 Å resolution where two-thirds of the rows, having indices of ($h - k \neq 3n$, $h + 2k \neq 3n$), are systematically weak. The other rows of strong reflections can be selected as those retaining an integral index after re-indexing *hkl* to ($h/3 - k/3$, $h/3 + 2k/3$, l), and they correspond to a small unit cell whose a, b-edges are reduced by a factor of $\sqrt{3}$.

Initial phasing in the small cell. Initial phases were obtained at 5.5 Å resolution by the MAD method[39] using a crystal labelled with Se-Met in the SmE–SmF–SmG proteins (SeEFG-MAD, Supplementary Table 1). The phases were calculated in the apparent space group of $P6_122$ and for the rows of strong reflections only. This procedure approximates both the rotational and translational ncs as crystallographic, and yields phases for the small unit cell ($a = 142.1$ Å, $c = 146.1$ Å for SeEFG-MAD) in which the asymmetric unit contains only one core domain representing the average of ncs copies with structural and orientational differences.

Fourteen Se sites were found using SHELXD[40] from the anomalous signal of the peak wavelength data of SeEFG-MAD. The Se positions were refined and phases calculated in SHARP[41]. The MAD phases showed an overall figure-of-merit of 0.649 (acentric) and 0.495 (centric) at 5.5 Å resolution for 12 sites. The anomalous phasing power was 1.742, 1.033 and 1.030 for the peak, inflection and remote wavelengths; the isomorphous phasing power was 0.913 (acentric) and 0.713 (centric) for the inflection, and 1.060 (acentric) and 0.860 (centric) for the remote. Using solvent-flattened phases from RESOLVE[42,43] and peak wavelength data to 5 Å resolution, the hetero-heptameric ring model for the core domain[16] was located in the density by a phased molecular replacement using MOLREP[44] in the automatic rotation-translation search mode. The search model consisted of a superposition of the ring model from ref. 16 and the homo-heptameric ring from the Lsm crystal structure (PDB ID 1M8V)[20]. The top solution showed Met residues in SmE–SmF–SmG proteins of the model overlapping with the majority of the Se anomalous peaks. The SeEFG-MAD phases were used to locate Se sites in the small cell, in anomalous difference maps calculated from the Se peak wavelength data from crystals containing Se-Met labels in other combinations of Sm proteins.

Partial model in the small cell. The SeEFG-MAD map of the small cell, even before phase improvement, showed clear density for an RNA helix later identified as part of the 3′-stem (Supplementary Fig. 10a). Presence for the bound Sm site RNA could be inferred from the oblong shape of the central hole of the ring. Density for the 5′-stem was poor but weakly connected to the Sm site region. There is density for the long L4 loop of SmD2, which was disordered in the SmD1–SmD2 heterodimer structure[16], but not for L4 of SmB, which was ordered in the SmD3–SmB heterodimer[16]. The regions showing poor densities at this stage were later found to exhibit greater variations among the ncs copies. A partial model containing the seven Sm proteins and a fragment of the 3′ RNA helix was built into the density in the small cell of the SeEFG-MAD derivative, using the program O[45] (Supplementary Fig. 10b).

Experimental map in the true cell. The partial model built in the small cell of SeEFG-MAD was used as the search model for molecular replacement by AMoRe[46] to the large cell (true cell) of each derivative in $P6_122$. Resolution for the rotation search was 7 Å; resolution for the translation searches was 4–6 Å for the first and second copies but reduced to 5–7.5 Å for the third copy as necessary. Acceptance of the translation solution was based on a minimal separation of 62 Å on finding the first copy, and displacement vectors between copies that agree with the self-Patterson. For all derivatives the solution after rigid-body refinement showed correlation with F_{obs} of >54% at 5.5 Å resolution. Thus the pseudo-translation was resolved.

The partial model for SeEFG-MAD was also mapped into the small cell of the other, non-isomorphous derivatives by molecular replacement in MOLREP[44]. For each derivative, the matrices relating the partial model in the small cell to its copies in the large cell were determined by the lsq_explicit command in O[45] and used to copy the Se sites into the large cell. Phases from different derivatives in the large cell were combined by multi-domain, multi-crystal averaging in DMMULTI[47] at 4.5 Å resolution, using three separate averaging masks over the ring, 3′- and 5′-stems of the partial model. Using the DMMULTI[47] phases with SeEFG-SAD amplitudes, we calculated an experimental map in the large $P6_122$ cell that allowed the building of the complete RNA.

Initial model in the large cell. The heptameric rings were rebuilt in the large cell. The Cα-trace of a homo-heptameric Lsm ring (PDB ID 1M8V)[20] was superimposed on the partial model by the Sm1, Sm2 motifs, to provide a regular scaffold for the seven inter-subunit β-sheets[16] that is a conserved characteristic of the Sm family. The crystal structures of SmD1, SmD2, SmD3 and SmB[16] were aligned to subunits in the 1M8V scaffold and adjusted to the density. For SmE, SmF and SmG, a composite template, consisting of the four Sm protein structures (PDB ID 1D3B, 1B34)[16], two Lsm structures (PDB ID 1LJO, 1M8V)[20,48], and the yeast SmF structure (PDB ID 1N9R)[49] in superposition, was aligned to the scaffold, and the model was rebuilt by modifying parts of the composite template that fitted the density better than the small-cell model. Additional adjustment was necessary to bring the Met side chains near the Se anomalous peaks. In SmF significant shifts of the β1 and β2 strands from the small-cell model were required to fit Se peaks for Met 20F and Met 27F.

Rigid-body refinement of sheet tilt. The model was subjected to refinement alternated with rebuilding using O[45] and Coot[50]. Starting from an *R*-factor of 56% at 4 Å resolution, rigid-body refinement was done in CNS[51] against the mlhl target of the DMMULTI[47] phases and SeEFG-SAD amplitudes, using one rigid group per core domain and then three groups of the ring, 5′- and 3′-stems. The resulting $2F_o - F_c$ map allowed the flanking RNA stems to be built in different conformations for the three ncs copies. Making each Sm protein a separate rigid group did not lower R_{free}, but making each inter-subunit β-sheet into a rigid group, involving cutting each protein across the middle of strands β2–β4, did reduce the R_{free} further by 2%. Therefore the seven inter-subunit β-sheets are found to have different tilts relative to the pseudo-sevenfold axis of the ring. Although the differences are moderate, the resulting architecture of the ring places the RNA binding loops (L3 and L5) in each Sm protein at different heights (Supplementary Fig. 8e, f) relative to the ring axis, thus predisposing the core ring to recognize RNA in circularly asymmetrical conformation.

Minimisation in $P6_122$. The model was further refined by minimisation in CNS at 3.6 Å resolution against the mlhl target of SeEFG-SAD amplitudes with DMMULTI[47] phases. The ncs restraints were applied in a single equivalence set that contained the Sm1, Sm2 motifs of all seven proteins and the Sm site RNA from the same ring. Additional restraints were imposed on base pairing in the RNA helices and protein main-chain hydrogen bond distances (2.9 ± 0.25 Å, 50 kcal mol^{-1} $Å^{-2}$). The R_{free} decreased to ~50%. Re-determining the Se sites using the model phase reduced the R_{free} of minimisation to ~48%, whereupon the improved $2F_o - F_c$ map allowed rebuilding of the Sm site region and the addition of the L4 loops of SmB as well as the N-, C-terminal extensions from the Sm-fold. When a fully Se-Met substituted derivative (Se-all, Supplementary Table 1) became available, the model built for the SeEFG-SAD crystal was positioned in that cell to calculate an

anomalous difference map at 5 Å resolution, which showed Se peaks accompanying each unique Se-Met residue in at least one copy. These peaks helped to place secondary structures in the extensions from the Sm fold. The SeEFG-SAD amplitudes were then replaced by the native, which reduced the R_{free} to ~46%.

Twinning. After re-indexing the native data in subgroups of $P6_122$, the self-rotation peaks at $\omega = 90^\circ$, $\kappa = 180^\circ$ fell with increasing resolution, indicating that the twofold rotation symmetries perpendicular to the *c*-axis are non-crystallographic. Analysis of the native intensities showed twinning about the *a,b*- and *a**,*b**-axes. The space group was re-classified as $P3_1$, and the model was duplicated by these twofold operators to generate six, then twelve copies per asymmetric unit, which were subjected to rigid-body refinement, giving R_{free} of 44.5%. *B*-factors were refined using data at 10–3.6 Å resolution. However, twin refinement by minimisation and bgroup using CNS, which assumes merohedral twinning, could not reduce R_{free} below 44%.

Twin refinement in $P3_1$. Refmac5 (refs 52, 53) found the crystal to be perfectly twinned with four twin domains of ~25% twin fractions. Correcting the representation of twinning immediately lowered R_{free} by ~10%. Rigid-body refinement with one core domain per rigid group against the anisotropy-corrected amplitudes at 20–5 Å reduced the R_{free} by 1.2% over 30 cycles. Further rigid-body refinement with one chain of protein or RNA per group caused some chains to rotate by more than 3° and reduced R_{free} by another 0.4%, to 32.6% at 5 Å resolution.

Three rounds of individual atom refinement in Refmac5 followed, with extensive rebuilding in-between. In the first round, 12-fold ncs restraints were still applied in a single equivalence set containing the Sm1, Sm2 motifs of all seven proteins and the Sm site RNA within one complex. In the later rounds, eight equivalence sets were defined, seven for the Sm1, Sm2 motifs of the seven proteins and one for the Sm site RNA. Main-chain hydrogen-bond restraints of 2.9 ± 0.2 Å were applied to the proteins throughout, using a list output by hydrogen_bonds.inp in CNS[51] and converted to the Refmac5 (ref. 52) format. No external restraints were imposed on the RNA. Ncs-averaged maps were calculated in Coot[50] from the $2F_o - F_c$ map, where the ncs matrices were evaluated between whole complexes using coordinates truncated to contain only the Sm1, Sm2 motifs and the Sm site RNA. The averaged map was used for rebuilding in the initial round wherever the $2F_o - F_c$ map was unclear; but in later rounds the $2F_o - F_c$ map was increasingly relied upon.

The first round of individual atom refinement was run for only one cycle to generate the $2F_o - F_c$ map at 3.6 Å resolution, which revealed that loop L1 of SmE, previously built to be like L1 of SmD2, was two residues too long and should be like L1 of the remaining Sm proteins. The other significant change was in the β-sheet of SmD1. Loop L1 forms the SmD1–SmD1 interface across an ncs twofold. After the rigid body refinement, at four of these interfaces there is a potential clash, which was relieved by shifting the β-strands towards β4 (see Supplementary Fig. 2) in accordance with the $2F_o - F_c$ map. This accounts for the greater root-mean-square Cα distances[54] among ncs copies of SmD1 compared with the other Sm proteins (Supplementary Table 2). In view of the modest resolution, Ramachandran restraints were turned on during real-space refinement in Coot.

The final model contains 7,285 protein, and 816 RNA residues. It has been refined at 66.2–3.6 Å resolution against 158,528 reflections (82.9% completeness) to $R = 27.6\%$, $R_{free} = 32.1\%$. An $F_o - F_c$ omit map of the Sm site, calculated by repeating the final refinement with A119–G125 omitted in all ncs copies, confirmed the conformation of the heptad (Supplementary Fig. 11). According to MolProbity[55], the model shows good stereochemistry in the 89th percentile of structures in the resolution range of 3.25–3.85 Å, with 91.8% of the protein residues in the favoured regions of the Ramachandran plot. All molecular structure figures were drawn using PyMOL[56].

31. Price, S. R., Ito, N., Oubridge, C., Avis, J. M. & Nagai, K. Crystallization of RNA–protein complexes. I. Methods for the large-scale preparation of RNA suitable for crystallographic studies. *J. Mol. Biol.* **249,** 398–408 (1995).
32. Sauter, C. *et al.* Additives for the crystallization of proteins and nucleic acids. *J. Cryst. Growth* **196,** 365–376 (1999).
33. Doublié, S. Preparation of selenomethionyl proteins for phase determination. *Methods Enzymol.* **276,** 523–530 (1997).
34. Leslie, A. G. W. The integration of macromolecular diffraction data. *Acta Crystallogr. D* **62,** 48–57 (2006).
35. Evans, P. R. Scaling and assessment of data quality. *Acta Crystallogr. D* **62,** 72–82 (2006).
36. French, G. S. & Wilson, K. S. On the treatment of negative intensity observations. *Acta Crystallogr. A* **34,** 517–525 (1978).
37. Kabsch, W. XDS. *Acta Crystallogr. D* **66,** 125–132 (2010) CrossRef.
38. Strong, M. *et al.* Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. USA* **103,** 8060–8065 (2006).
39. Hendrickson, W. A. & Ogata, C. M. Phase determination from multiwavelength anomalous diffraction measurements. *Methods Enzymol.* **276,** 494–523 (1997).
40. Schneider, T. R. & Sheldrick, G. M. Substructure solution with SHELXD. *Acta Crystallogr. D* **58,** 1772–1779 (2002).
41. de La Fortelle, E. & Bricogne, G. Maximum likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol.* **276,** 472–494 (1997).
42. Terwilliger, T. C. SOLVE and RESOLVE: automated structure solution and density modification. *Methods Enzymol.* **374,** 22–37 (2003).
43. Terwilliger, T. C. SOLVE and RESOLVE: automated structure solution, density modification and model building. *J. Synchrotron Radiat.* **11,** 49–52 (2004).
44. Vagin, A. & Teplyakov, A. MOLREP: an automated program for molecular replacement. *J. Appl. Cryst.* **30,** 1022–1025 (1997).
45. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).
46. Navaza, J. Implementation of molecular replacement in AmoRe. *Acta Crystallogr. D* **57,** 1367–1372 (2001).
47. Cowtan, K. D., Zhang, K. Y. J. & Main, P. In *International Tables for Crystallography, Volume F. Crystallography of Biological Macromolecules* (eds. Rossmann, M. G. & Arnold, E.) 705–710 (Kluwer Academic Publishers, 2001).
48. Törö, I., Basquin, J., Teo-Dreher, H. & Suck, D. Archaeal Sm proteins form heptameric and hexameric complexes: crystal structures of the Sm1 and Sm2 proteins from the hyperthermophile *Archaeoglobus fulgidus*. *J. Mol. Biol.* **320,** 129–142 (2002).
49. Collins, B. M. *et al.* Homomeric ring assemblies of eukaryotic Sm proteins have affinity for both RNA and DNA. Crystal structure of an oligomeric complex of yeast SmF. *J. Biol. Chem.* **278,** 17291–17298 (2003).
50. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of *Coot*. *Acta Crystallogr. D* **66,** 486–501 (2010).
51. Brunger, A. T. Version 1.2 of the Crystallography and NMR system. *Nature Protocols* **2,** 2728–2733 (2007).
52. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53,** 240–255 (1997).
53. Vagin, A. *et al.* Organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr. D* **60,** 2184–2195 (2004).
54. Diamond, R. On the multiple simultaneous superposition of molecular structures by rigid body transformations. *Protein Sci.* **1,** 1279–1287 (1992).
55. Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35,** W375–W383 (2007).
56. DeLano, W. L. The PyMOL Molecular Graphics System ⟨http://www.pymol.org⟩ (2002).

LETTER

doi:10.1038/nature09964

Improved molecular replacement by density- and energy-guided protein structure optimization

Frank DiMaio[1], Thomas C. Terwilliger[2], Randy J. Read[3], Alexander Wlodawer[4], Gustav Oberdorfer[5], Ulrike Wagner[5], Eugene Valkov[6], Assaf Alon[7], Deborah Fass[7], Herbert L. Axelrod[8], Debanu Das[8], Sergey M. Vorobiev[9], Hideo Iwaï[10], P. Raj Pokkuluri[11] & David Baker[1]

Molecular replacement[1–4] procedures, which search for placements of a starting model within the crystallographic unit cell that best account for the measured diffraction amplitudes, followed by automatic chain tracing methods[5–8], have allowed the rapid solution of large numbers of protein crystal structures. Despite extensive work[9–14], molecular replacement or the subsequent rebuilding usually fail with more divergent starting models based on remote homologues with less than 30% sequence identity. Here we show that this limitation can be substantially reduced by combining algorithms for protein structure modelling with those developed for crystallographic structure determination. An approach integrating Rosetta structure modelling with Autobuild chain tracing yielded high-resolution structures for 8 of 13 X-ray diffraction data sets that could not be solved in the laboratories of expert crystallographers and that remained unsolved after application of an extensive array of alternative approaches. We estimate that the new method should allow rapid structure determination without experimental phase information for over half the cases where current methods fail, given diffraction data sets of better than 3.2 Å resolution, four or fewer copies in the asymmetric unit, and the availability of structures of homologous proteins with >20% sequence identity.

The limiting steps in molecular replacement are finding the correct location of the starting model in the unit cell and the interpretation of electron density maps produced using the imperfect phase information from candidate model placements. The left column of Fig. 1 illustrates the problem of initial model-building starting with distant comparative models (20–30% sequence identity) that have been correctly placed in the crystallographic unit cell. Automatic chain tracing methods fail on such maps because they often follow the incorrect comparative model (red) more closely than the actual structure (yellow); breaks in the density make it difficult to recover the correct backbone trace. Nevertheless, the maps contain considerable information about the native structure; for example, portions of the starting model that are not within density are generally incorrect.

Structure prediction methods such as Rosetta search for the lowest energy conformation of the polypeptide chain using physically realistic force fields. Based on previous work showing that accurate structures could be obtained from very sparse NMR data sets[15] by using the data to guide structure prediction searches, we reasoned that structure prediction methods guided by even very noisy density maps might be able to improve a poor molecular replacement model before applying crystallographic model-building techniques. We developed an approach in which electron density maps generated from molecular replacement solutions for each of a series of starting models are used to guide energy optimization by structure rebuilding, combinatorial side chain packing, and torsion space minimization[16]. New maps are generated using phase information from the energy-optimized models most consistent with the diffraction data, subjected to automatic chain tracing, and success is monitored through the free R factor[17].

To investigate the performance of the new method, we obtained 18 crystallographic data sets that had resisted previous attempts at structure determination. We first tested whether a comprehensive set of state-of-the-art molecular replacement approaches using a range of full-length and trimmed templates and homology models could solve any of these structures (Supplementary Information). We were able to solve five of the structures with both the new method and the existing methods (Table 1), leaving 13 challenging data sets highly resistant (Supplementary Information section 1) to structure determination (Table 1). For each of these, we identified homologous proteins of known structure[18] and constructed sequence alignments and starting models[9] from the five closest homologues. Starting models were used to search for up to five

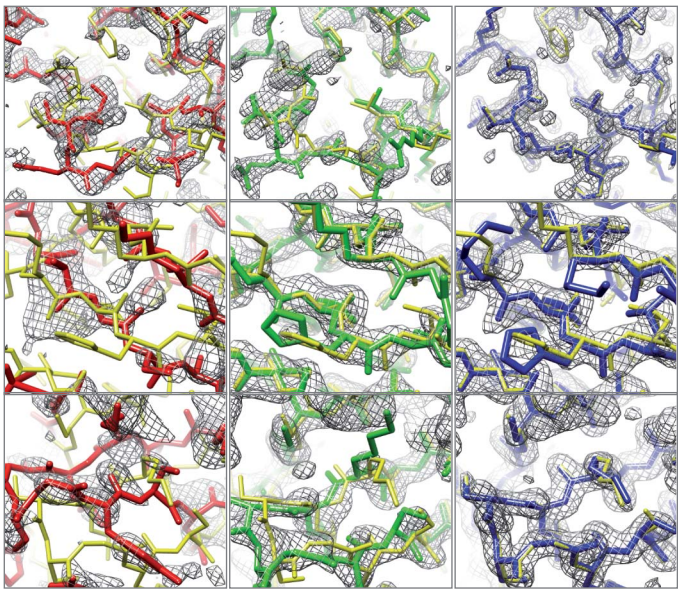


Figure 1 | Examples of improvement in electron density and model quality. Each row corresponds to one of the entries in Table 1. First row: 6 (2.0 Å resolution); second row: 7 (2.1 Å resolution); third row: 12 (1.7 Å resolution). Left column: correct initial molecular replacement solution (not necessarily identifiable at this stage) using starting model and corresponding density. Middle column: energy-optimized model and corresponding density. Right column: model and density following automatic building using the energy-optimized model as the source of phase information. The final deposited structure is shown in yellow in each panel; the initial model, energy-optimized model, and model after chain rebuilding are in red, green and blue, respectively. The sigma-A-weighted $2mF_o - DF_c$ density contoured at 1.5σ is shown in grey.

[1]University of Washington, Department of Biochemistry and HHMI, Seattle, Washington 98195, USA. [2]Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA. [3]University of Cambridge, Department of Haematology, Cambridge Institute for Medical Research, Cambridge CB2 0XY, UK. [4]Macromolecular Crystallography Laboratory, National Cancer Institute at Frederick, Frederick, Maryland 21702, USA. [5]Institute of Molecular Biosciences, University of Graz, Humboldtstrasse 50/3, 8010-Graz, Austria. [6]University of Cambridge, Department of Biochemistry, Cambridge CB2 1GA, UK. [7]Weizmann Institute of Science, Department of Structural Biology, Rehovot 76100, Israel. [8]Joint Center for Structural Genomics and SSRL, SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA. [9]Northeast Structural Genomics Consortium, Columbia University, New York, New York 10027, USA. [10]University of Helsinki, Institute of Biotechnology, FI-00014 Helsinki, Finland. [11]Argonne National Laboratory, Biosciences Division, Argonne, Illinois 60439, USA.

Table 1 | Determination of previously unsolved structures using the new approach

ID number	Source*	Resolution (Å)	Seqid (%)	R_{free} after Phaser MR and model-building protocol							
				Autobuild	Arp/Warp	Simulated annealing (SA) + Autobuild	Torsion-space SA + Autobuild	Extreme SA + Autobuild	DEN + Autobuild	Rosetta + Autobuild	R_{free} (current best)
Solved by multiple methods											
1	JCSG	2.1	22	**0.31**	0.50	0.30	**0.30**	**0.30**†	**0.35**	**0.31**	0.22
2	NSGC	2.2	19	**0.29**	0.57	0.29	**0.29**	**0.29**†	**0.30**	**0.29**	0.22
3	UG	2.5	27	**0.34**	0.59	0.29	**0.29**	**0.29**†	**0.35**	**0.27**	0.19
4	JCSG	2.7	21	**0.31**	0.59	0.30	**0.30**	**0.30**†	**0.31**	**0.30**	0.24
5	ANL	1.9	31	0.51	0.59	0.54	0.54	**0.24**	**0.39**	**0.31**	0.24
Only solved by Rosetta											
Rosetta modelling with density required for successful model-building											
6	NCI	2.0	30	0.56	0.59	0.60	0.55	0.55	0.50	**0.34**	0.20
7	WI	2.1	22/15	0.56	0.60	0.54	0.54	0.54	0.56	**0.28**	0.26
8	JCSG	2.8	29	0.52	0.55	0.50	0.50	0.51	0.45	**0.36**	0.36‡
9	UC	3.0	22	0.54	0.56	0.50	0.50	0.47	0.46	**0.32**	0.25§
10	JCSG	3.2	20	0.54	0.57	0.51	0.51	0.53	0.46	**0.39**	0.33‡
11	UG	2.5	18	0.52	0.57	0.54	0.52	0.54	0.55	**0.27**	0.22
	MEAN			0.54	0.57	0.53	0.52	0.52	0.50	**0.33**	
Rosetta homology modelling required for successful molecular replacement											
12	BI,HY	1.7	– (100)	–	–	–	–	–	–	**0.29**	0.22
13‖	JCSG	2.9	29	–	–	–	–	–	–	**0.39**	0.23

The Seqid column gives the sequence identity to the closest homologue identified by HHpred[18], and is shown in parentheses if this is an NMR structure. The next seven columns give the R_{free} of the model produced by different combinations of refinement and autobuilding approaches. The final column gives the R_{free} after further refinement by the crystallographer who provided the data. For structures solved by multiple methods, the new method as well as one or more alternative approaches was sufficient ($R_{free} < 0.4$). In the first subset of structures that could only be solved by the new method (only solved by Rosetta), molecular replacement succeeds (in some cases ambiguously) using the template alone but model-building fails; in the second subset, refinement in Rosetta is required for molecular replacement to succeed. Targets that could not be solved by our approach are listed in Supplementary Table 1.

* JCSG, Joint Center for Structural Genomics; NSCG, Northeast Center for Structural Genomics; UG, University of Graz; ANL, Argonne National Lab; NCI, National Cancer Institute; WI, Weizmann Institute of Science; UC, University of Cambridge; BI,HY, Institute Of Biotechnology, University of Helsinki.

† Because a single SA trajectory was sufficient to solve these cases, Extreme SA was not run. Values from the single SA run are shown for completeness.

‡ Solutions for both are essentially correct based on the selenium positions in the anomalous difference Fourier maps calculated from the experimental data. However, structures are difficult to complete to deposition due to some MR solution model bias, poor or disordered density in numerous regions and low resolution.

§ Refinement ongoing.

‖ This structure was solved and all tests on this template were carried out using the intact template as a starting point. With this template both the molecular replacement step and subsequent rebuilding required Rosetta modelling for success. After determining the structure and completing the tests we found that it was also possible to solve the structure by molecular replacement if the template were split into two rigid subunits and the two domains were correctly chosen.

candidate molecular replacement solutions based on the likelihood of the experimental diffraction data[2]. Electron density maps were computed for each of these solutions, and used to guide energy minimization by first remodelling the unaligned regions and regions which poorly fit the density and then optimizing all backbone and side chain torsion angles. The likelihood of the experimental diffraction data was computed for each optimized model[2]; if top ranked models were similar (see Methods), a map generated from the highest likelihood model was subjected to automatic chain rebuilding, density modification and refinement[5]. If this succeeded in building the majority of the protein and produced a model with free *R* factor[17] significantly better than random ($R_{free} < 0.4$), the structure was considered solved; rebuilt models were further analysed by the crystallographers who supplied the original data. Using this approach, we were able to solve eight of the thirteen challenging cases (Table 1). In some of these eight cases, recognition of the correct placement of the model in the unit cell was only possible after Rosetta refinement (Supplementary Fig. 2); in others the correct placement was clear but the density was too poor for chain rebuilding. In two of the cases (12 and 13), even finding the correct molecular replacement solution first required energy-based refinement[12].

The improvement in electron density produced by density guided energy optimization and autobuilding are illustrated in Fig. 1. The starting molecular replacement models are often quite inaccurate, and the density generated from these models has breaks within the backbone of the actual structure (left panels). After model rebuilding and energy guided structure optimization, backbone breaks are largely closed and both side chains and backbone are more correctly modelled (middle panels). Automatic chain rebuilding into the improved map followed by density modification and reciprocal-space refinement further improve the model and the density (right panels). For all eight cases, the correlation between the final refined density and density from the original molecular replacement solutions is low, increases significantly after energy- and density-based structure optimization, and still further after automatic chain rebuilding (Supplementary Table 2).

For each of the eight challenging cases solved with the new method we also applied a battery of existing methods (Table 1 and Supplementary Information section 1) including simulated annealing in Cartesian and torsion space in PHENIX and CNS[14], deformable elastic network (DEN) refinement[13] in CNS, and PHENIX Autobuild[6] and ARP-WARP[5] for model-building. As noted above, in two cases Rosetta structure modelling was required for the correct placement of starting models in the unit cell, so the alternative methods could not even be applied. In the remaining six cases, final R_{free} values were lower using the new approach than with any of the existing methods (Table 1, Fig. 2a). Whereas conventional simulated annealing in both Cartesian and torsion space had little effect, the recently developed DEN[19] refinement protocol did improve three of the structures slightly, yielding free *R* values of 0.45–0.46 for these targets. Combination of DEN refinement with the method described here could lead to still more powerful approaches.

To benchmark the sequence and structural divergence where the different methods break down, we studied two different protein families for which a total of 59 different template structures covering a broad range of sequence and structural similarity were available (Supplementary Tables 3–5). Each template was correctly placed in the unit cell, and then improved with either Rosetta energy- and density-based optimization, Cartesian- and torsion-space simulated annealing, or DEN refinement. For each resulting model, the correlation with the density of the deposited structure was evaluated. Automatic chain rebuilding beginning with the superimposed starting models was successful for 18 of the 59 cases, consistent with the observation that molecular replacement often fails with templates sharing less than 30% sequence identity with the target sequence. Torsion-space simulated annealing in CNS before autobuilding allowed solution of two additional structures, DEN refinement, three additional structures, and Rosetta energy-based structure optimization, fourteen additional structures (Supplementary Fig. 2 and Supplementary Tables 3–5). We found the radius of convergence of the new method can be further extended by guiding energy based structure optimization by the Patterson correlation[20] rather than electron density (see Supplementary Information). This allowed structure improvement and identification of the correct molecular replacement solution in two additional cases (Supplementary Fig. 2, compare green

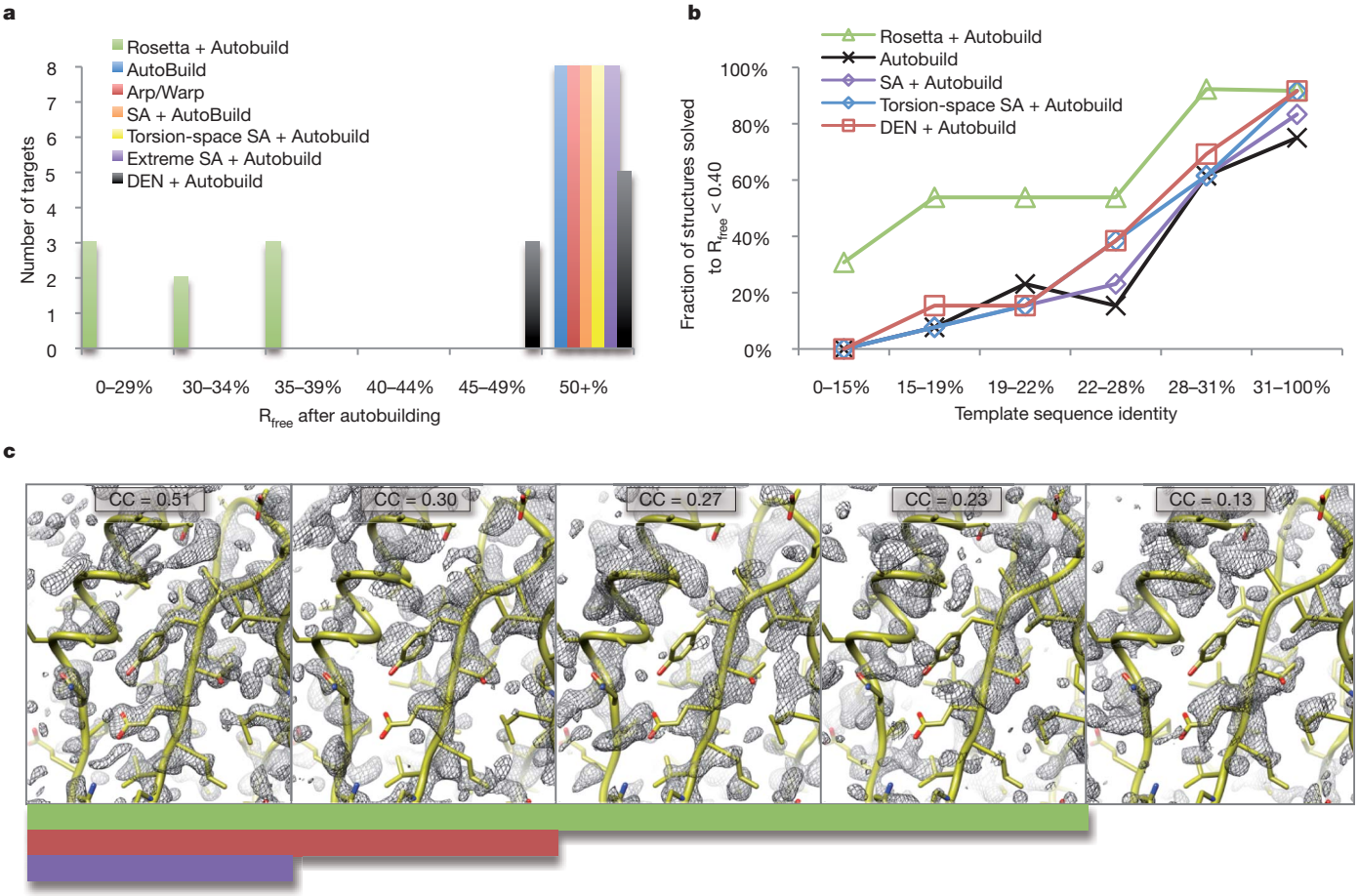


Figure 2 | Method comparison. **a**, Histogram of R_{free} values after autobuilding for the eight difficult blind cases solved using the new approach (Table 1). For most existing approaches, none of the cases yielded R_{free} values under 50%; DEN was able to reduce R_{free} to 45–49% for three of the structures. For all eight cases, Rosetta energy and density guided structure optimization led to R_{free} values under 40%. **b**, Dependence of success on sequence identity. The fraction of cases solved (R_{free} after autobuilding <40%) is shown as a function of template sequence identity over the 18 blind cases and 59 benchmark cases. The new method is a clear improvement below 28% sequence identity. **c**, Dependence of structure determination success on initial map quality. Sigma-A-weighted $2mF_o - DF_c$ density maps (contoured at 1.5σ) computed from benchmark set templates with divergence from the native structure increasing from left to right are shown in grey; the solved crystal structure is shown in yellow. The correlation with the native density is shown above each panel. The solid green bar indicates structures the new approach was able to solve ($R_{free} < 0.4$); the red bar those that torsion-space refinement or DEN refinement is able to solve, and the purple bar those that can be solved directly using the template.

to orange bar); for one of these the improvements were sufficient for autobuilding to effectively solve the structure.

Over the combined set of 18 blind cases and the 59 benchmark cases, Rosetta refinement yielded a model with density correlation as good or better than any of the control methods for all but six structures. The dependence of success on sequence identity over the combined set is illustrated in Fig. 2b. The improvement in performance is particularly striking below 22% sequence identity, where the quality of the starting homology models becomes too low for the control methods in almost all cases. With the new method the success rate in the 15–28% sequence identity range, generally considered very challenging for molecular replacement, is over 50%.

Figure 2c illustrates the dependence of model-building on the quality of initial electron density. Conventional chain rebuilding requires a map in which the connectivity is largely correct (leftmost panel), whereas the new method can tolerate breaks in the chain more than other methods (panels 2–4), as long as there is sufficient information in the electron density map, combined with the Rosetta energy function, to guide structure optimization. The map on the far right contains too little information to guide energy-based structure optimization and hence the new approach fails. In the five blind cases that have not yet been solved the comparative models may have been too low in quality, or there may have been complications in the X-ray diffraction data sets themselves.

Key to the success of the approach described here is the integration of structure prediction and crystallographic chain tracing and refinement methods. Simulated annealing guided by molecular force fields and diffraction data has had an important role in crystallographic refinement[14,21]. Structure prediction methods such as Rosetta can be even more powerful when combined with crystallographic data because the force fields incorporate additional contributions such as solvation energy and hydrogen bonding, and the sampling algorithms can build non-modelled portions of the molecule *de novo* and cover a larger region of conformational space than simulated annealing. The difference between Rosetta sampling and simulated annealing sampling, both using crystallographic data, is illustrated in Fig. 3. Beginning with the homology model placed by molecular replacement in the unit cell for blind case 6, we generated 100 models by simulated annealing at two starting temperatures, and 100 models with Rosetta energy- and density-guided optimization followed by refinement. The $2mF_o - DF_c$ (ref. 22) electron density maps generated using phases from over 50% of the Rosetta models had correlations 0.36 or better to the final refined map, whereas fewer than 5% of models from simulated annealing had correlations this high. Our approach probably outperforms even extreme simulated annealing because the physical chemistry and protein structural information which guide sampling eliminate the vast majority of non-physical conformations.

Approaches to molecular replacement combining the power of crystallographic map interpretation and structure prediction methodology

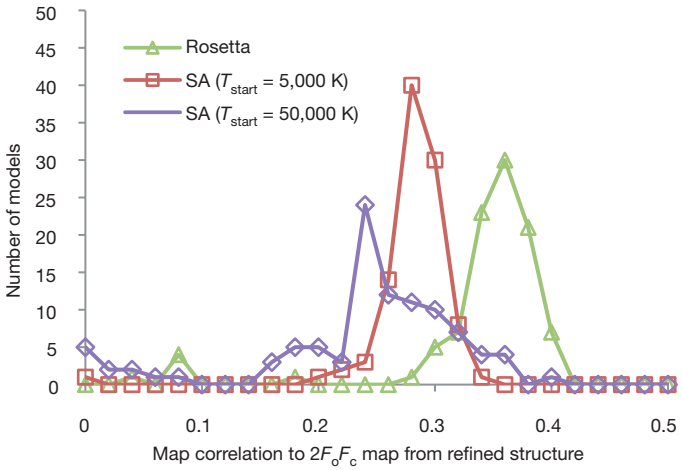


Figure 3 | Comparison of the effectiveness of model diversification using Rosetta and simulated annealing. For blind case 6, 100 models were generated using either simulated annealing with a start temperature of 5,000 K, simulated annealing with a start temperature of 50,000 K, or Rosetta energy- and density-guided optimization. The correlation between $2mF_o - DF_c$ density maps computed from each structure and the final refined density was then computed; the starting model has a correlation of 0.29 and the distributions of the refined models are shown in the figure. Rosetta models have correlations better than the initial model much more often than simulated annealing.

are likely to become increasingly useful in the next few years. First, the number of already-determined structures will continue increasing, making it increasingly likely that there will be a structure with the required >20% sequence identity: the chance there is a structure with a sequence identity of 20% or greater is more than twice that of finding a structure with at least 30% sequence identity[23]. Second, as more work focuses on proteins that cannot be expressed in *Escherichia coli*, the currently preferred methods for experimental phase determination based on selenomethionine replacement may be more difficult to apply. Finally, as protein structure modelling algorithms improve, better initial models should further increase the radius of convergence of the approach.

METHODS SUMMARY

Starting models (templates) for molecular replacement were generated by searching the PDB using HHpred[18] for proteins likely to have structures related to the query. Starting models were constructed from alignments generated by HHpred. Unaligned residues were removed from the template and non-identical side chains were stripped back to the gamma carbon (CG), as suggested in previous work[9]. An initial Phaser search with a low rotation function cutoff (50%) and modest packing threshold (up to 10 clashes) was used to find up to five putative molecular replacement (MR) solutions for each template. Each MR solution for each template was used to obtain an initial estimate of phases and the corresponding sigma-A-weighted $2mF_o - DF_c$ density map was generated[22]. Gaps in the initial alignment, as well as regions around deletions, were rebuilt using the Rosetta loop modelling protocol[12], which alternates insertion of short fragments with similar local sequences and cyclic coordinate descent (CCD) closure[24]. Twenty-four rounds of side chain rotamer optimization and side chain and backbone torsion-space minimization were then used to optimize a linear combination of the Rosetta all-atom energy and a term assessing agreement to the electron density. Following the energy- and density-guided refinement, models were ranked based on the Phaser log-likelihood score. The highest ranked models were then subjected to a second round of modelling using the Rosetta iterative rebuild and refine protocol[12] constrained by density. After this final round of refinement, the model with best agreement to the experimental data (highest likelihood) was used to either find additional models in the asymmetric unit, or as a starting point for Phenix AutoBuild.

The procedures described here require considerable computation as up to several thousand Rosetta models are generated for each structure, typically requiring 0.5–1 h per structure of CPU time. We have developed automated procedures in Phenix (*phenix.mr_rosetta*) that use Rosetta and Phenix modules to carry out and extend many of the methods described here with density modification and density averaging, potentially allowing fewer Rosetta models to be used. All the methods described in this paper are available in release 3.2 of Rosetta.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 4 August 2010; accepted 22 February 2011.
Published online 1 May 2011.

1. Rossmann, M. G. *The Molecular Replacement Method* (Gordon & Breach, 1972).
2. McCoy, A. J. *et al. Phaser* crystallographic software. *J. Appl. Cryst.* **40,** 658–674 (2007).
3. Brünger, A. T. *et al. Crystallography & NMR system*: a new software system for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
4. Vagin, A. & Teplyakov, A. *MOLREP*: an automated program for molecular replacement. *J. Appl. Cryst.* **30,** 1022–1025 (1997).
5. Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nature Protocols* **3,** 1171–1179 (2008).
6. Terwilliger, T. C. *et al.* Iterative model building, structure refinement and density modification with the *PHENIX AutoBuild* wizard. *Acta Crystallogr. D* **64,** 61–69 (2008).
7. DePristo, M. A., de Bakker, P. I. W., Johnson, R. J. K. & Blundell, T. L. Crystallographic refinement by knowledge-based exploration of complex energy landscapes. *Structure* **13,** 1311–1319 (2005).
8. Cowtan, K. The *Buccaneer* software for automated model building. *Acta Crystallogr. D* **62,** 1002–1011 (2006).
9. Schwarzenbacher, R., Godzik, A., Grzechnik, S. K. & Jaroszewski, L. The importance of alignment accuracy for molecular replacement. *Acta Crystallogr. D* **60,** 1229–1236 (2004).
10. Rodríguez, D. D. *et al.* Crystallographic *ab initio* protein structure solution below atomic resolution. *Nature Methods* **6,** 651–653 (2009).
11. Suhre, K. & Sanejouand, Y. H. On the potential of normal-mode analysis for solving difficult molecular-replacement problems. *Acta Crystallogr. D* **60,** 796–799 (2004).
12. Qian, B. *et al.* High-resolution structure prediction and the crystallographic phase problem. *Nature* **450,** 259–264 (2007).
13. Schröder, G., Levitt, M. & Brünger, A. T. Super-resolution biomolecular crystallography with low-resolution data. *Nature* **464,** 1218–1222 (2010).
14. Brünger, A. T., Kuriyan, J. & Karplus, M. Crystallographic *R* factor refinement by molecular dynamics. *Science* **235,** 458–460 (1987).
15. Raman, S. *et al.* NMR structure determination for larger proteins using backbone-only data. *Science* **327,** 1014–1018 (2010).
16. Das, R. & Baker, D. Macromolecular modeling with Rosetta. *Annu. Rev. Biochem.* **77,** 363–382 (2008).
17. Brünger, A. T. Free *R* value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355,** 472–475 (1992).
18. Söding, J. Protein homology detection by HMM–HMM comparison. *Bioinformatics* **21,** 951–960 (2005).
19. Schröder, G. F., Brunger, A. T. & Levitt, M. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure* **15,** 1630–1641 (2007).
20. Brünger, A. T. Extension of molecular replacement: a new search strategy based on Patterson correlation refinement. *Acta Crystallogr. A* **46,** 46–57 (1990).
21. Brünger, A. T., Karplus, M. & Petsko, G. A. Crystallographic refinement by simulated annealing: application to crambin. *Acta Crystallogr. A* **45,** 50–61 (1989).
22. Read, R. J. Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr. A* **42,** 140–149 (1986).
23. Vitkup, D., Melamud, E., Moult, J. & Sander, C. Completeness in structural genomics. *Nature Struct. Biol.* **8,** 559–566 (2001).
24. Canutescu, A. & Dunbrack, R. Cyclic coordinate descent: a new algorithm for loop closure in protein modeling. *Protein Sci.* **12,** 963–972 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements R.J.R., T.C.T. and D.B. thank the NIH (5R01GM092802), the Wellcome Trust (R.J.R.), and HHMI (D.B.) for funding this research. F.D. acknowledges the NIH (P41RR002250) and HHMI. D.F. and A.A. acknowledge support from the Israel Science Foundation. G.O. thanks DK Molecular Enzymology (FWF-project W901) and the Austrian Science Fund (FWF-project P19858). The work of A.W. was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. H.I. acknowledges support from the academy of Finland (1131413). S.M.V. was supported by a grant from the Protein Structure Initiative of National Institute of General Medical Sciences (U54 GM074958). The work of P.R.P. at Argonne National Laboratory was supported by the US Department of Energy's Office of Science, Biological and Environmental Research GTL programme under contract DE-AC02-06CH11357. We thank all members of the JCSG for their general contributions to the protein production and structural work. The JCSG is supported by the NIH, National Institutes of General Medical Sciences, Protein Structure Initiative (U54 GM094586 and GM074898).

Author Contributions F.D., T.C.T., R.J.R. and D.B. developed the methods described in the manuscript; F.D., T.C.T., R.J.R., A.W. and D.B wrote the paper. A.W., G.O., U.W., E.V., A.A., D.F., H.L.A., D.D., S.M.V., H.I. and P.R.P. provided the data and refined one or more structures to completion.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to T.C.T. (terwilliger@lanl.gov) or D.B. (dabaker@u.washington.edu).

METHODS

Preparation of templates and identification of initial molecular replacement solutions. For the application of the new method to blind cases, templates were identified using HHpred[18]. For both the blind and benchmark data sets, HHpred was used to generate initial alignments. We prepared templates by removing all unaligned residues and stripping all non-identical side chains to the gamma carbon (CG), as suggested in previous work[9]. An initial Phaser search with a low rotation function cutoff (50%) and modest packing threshold (up to 10 clashes) was used to find up to five putative MR solutions for each template. In two blind cases (12 and 13 in Table 1), Phaser was unable to locate the correct configuration of a molecule using the template alone, but modelling in Rosetta without density-fitting constraints before Phaser search enabled discovery of the correct rigid-body placement of the molecule, with very low Phaser translation function *Z*-scores (TFZ) of 4–6 (after solving 13, it was discovered that breaking the template into two rigid subunits enabled solution of the molecular replacement problem). If point-group symmetry was present in the templates, the initial search (and subsequent steps) were carried out both with monomeric and multimeric models (see subsection on symmetric modelling into density below).

Rebuilding and refinement into density. Each MR solution for each template was used to obtain an initial estimate of phases and the corresponding sigma-A-weighted $2mF_o - DF_c$ density map was generated[22]. Gaps in the initial alignment, as well as regions around deletions, were rebuilt using the Rosetta loop modelling protocol[12], which alternates insertion of short fragments with similar local sequences and CCD closure[24]. Twenty-four rounds of side chain rotamer optimization and side chain and backbone torsion-space minimization were then used to optimize a linear combination of the Rosetta all-atom energy and a term assessing agreement to the electron density. Agreement to density was computed using an extension of a method previously developed for building into cryo-electron microscopy density[25]. Density was calculated from a model using a single-Gaussian approximation to atomic scattering factors. Correlation coefficients between model and map were calculated for each residue: the computed density includes all atoms in the residue and the backbone in the two flanking residues on each side, and the correlation is taken over a mask extending 5 Å from each atom. Scores are proportional to the negative log probability that observed correlations occur by random chance, assuming a normal distribution; parameters are trained matching randomly oriented fragments into synthesized density. In all cases, density was truncated at 3 Å.

Following the energy- and density-guided refinement, models were ranked based on the Phaser log-likelihood score. The highest ranked models were then subjected to a second round of modelling using the Rosetta iterative rebuild and refine protocol[12] constrained by density. Regions that deviated the most from the current estimate of the electron density were rebuilt; clashes between crystallographic (and non-crystallographic) contacts were also always rebuilt. For each template carried over to the second round (typically the top-scoring 3–10 models from the previous round), 2,000 Rosetta models were generated. The likelihood of the diffraction data was again computed using Phaser for the lowest-energy 10% of models, and if the five highest likelihood models were in the same rigid-body configuration (that is if they had density correlations above 0.2 with each other), they were used to re-phase the density and an additional round (24 cycles) of side chain optimization and refinement was carried out in Rosetta. If the top-scoring models differed, then additional templates were considered (if available) or Rosetta homology modelling was used to perturb the initial structures before molecular replacement.

After this final round of refinement, the model with best agreement to the experimental data (highest likelihood) was used to either find additional models in the asymmetric unit, or as a starting point for Phenix AutoBuild. In cases where the R_{free} was better than random but higher than 0.4, and a majority of residues were placed, additional refinement was carried out using models produced by AutoBuild, which allows for recovery from sequence alignment errors. The bond lengths and bond angles were first replaced with ideal values with small compensating changes in the torsion angles to minimize the change in interatomic distances, and the idealized models were then subjected to 48 cycles of side chain rotamer optimization and side chain and backbone torsion minimization. In the first 24 cycles, the Rosetta all-atom energy function was optimized, and in the final 24 cycles a weighted sum of Rosetta all-atom energy and the fit-to-density energy described above was optimized.

Refinement of symmetric complexes into density. Key to solving many of the blind cases was proper treatment of symmetry. In cases where there is point-group symmetry in the asymmetric unit (either from the template or subsequently discovered by molecular replacement search) or there is close contact between crystal partners, the Rosetta symmetric modelling framework[26] was used to reduce the size of the conformational space which must be searched. This occurred in blind cases where either there was point-group symmetry in the template(s) (6 in Table 1), point-group symmetry was found during the Phaser search (13), or tight crystal contacts formed point-group symmetry (8 and 10). In these cases, Rosetta optimizes only the torsion angles in one subunit and the rigid-body degrees of freedom of the corresponding symmetric group. The energy is calculated explicitly over a non-redundant subset of atoms for computational efficiency, but the fit to density is calculated without symmetrization. This is similar to the "strict formulation" of symmetry introduced in ref. 27.

Symmetric modelling in Rosetta requires that the energy of a symmetric complex be expressible in terms of a single subunit or as pairwise interactions between this subunit and other ones. Minimization also only considers gradients from these components. To take advantage of Rosetta's symmetric modelling with asymmetric density data, the gradients of each subunit with respect to the fit-to-density energy must be mapped to a single subunit. The score of a residue *i*'s fit to density is just the sum of the fit-to-density scores over all of *i*'s copies. As a first approximation, the gradient at *i* can be computed as the combined gradients of all of *i*'s copies, rotated by the symmetry operation to rotate the subunit containing *i*'s copy to the one containing *i*. Unfortunately, although this approach correctly handles gradients of internal torsions, the gradients at each symmetric degree-of-freedom are not correctly handled. Proper handling takes advantage of the formulation from ref. 28 to efficiently convert Cartesian gradients to torsion-space gradients. For each atom in the symmetric complex, we compute F_1 and F_2 corresponding to the unrotated gradient with respect to the fit-to-density score. For internal torsional degrees of freedom, the rotation applied to each F_1/F_2 just maps each subunit back to the asymmetric unit. At each symmetric degree of freedom we apply a corresponding symmetry operation; for example, in D3 symmetry (a dimer of trimers) the degree of freedom corresponding to the "spin" of the trimers applies the rotation used to transform between trimers to all the F_1/F_2's in one of the trimers.

Refinement against the Patterson function. In benchmark cases where the Phaser translation search failed to find the correct molecular placement even when many potential solutions were considered, we conducted refinement against the Patterson function. A score function was implemented that assessed the correlation between the computed and experimental Patterson map (next paragraph). The map was truncated to between 3.5 Å and 10 Å resolution (in reciprocal space) and 5 Å to ~75% of the template diameter (in real space). Starting models used the same templates and rebuilding procedure as the density refinement. Because the correct rotation is not known at this stage, the molecule orientation was randomized at the beginning of each refinement trajectory and constraints on backbone atoms were used to prevent the molecule from rotating more than ~5° from this starting orientation.

The scoring function we optimize is the weighted sum of Rosetta's all-atom potential function and the correlation between the calculated Patterson map and the observed Patterson map. To make this tractable in Rosetta refinement, which may require tens of thousands of score-function evaluations per trajectory, simplifications are necessary. Directly computing $\partial p_{calc}/\partial x$ requires three fast Fourier transforms (FFTs) per atom. However, since what is needed is not $\partial p_{calc}/\partial x$ but instead the sum $\partial \sum_{map} p_{calc}p_{obs}/\partial x$, FFTs can be used to compute the change in correlation at every position in the map at once (where p is the Patterson density and ρ is the real-space density):

$$\frac{\partial \sum_{map} p_{calc}p_{obs}/\partial x}{\partial x} = F^{-1}[F[p_{obs}]\cdot F[\rho_{calc}]\cdot[F[\partial \rho_{calc}/\partial x_i]](x) \quad (1)$$

Assuming a fixed B-factor over the molecule, this requires just 3 FFTs per atom type (the correction terms that make this not just the overlap integral but a true correlation can be folded into the same FFT). Then, given a model to refine against the Patterson map, we compute equation (1) once, sum over all the symmetric orientations of the space group, and interpolate the gradient at each atom's position. Given sufficiently fine sampling, this gives a very close approximation to the true derivative in a small fraction of the CPU time.

For side chain optimization, where we must rescore the Patterson correlation for exponentially many combinations of side chain rotamers, exact computation is also intractable. However, first computing the density ρ_{calc} of the backbone only, then computing the correlation scores for each side chain rotamer independently, provides a reasonably good approximation with only several hundred to several thousand function evaluations (one for each rotamer).

Torsion space simulated annealing with DEN restraints. As a control, we ran torsion-space simulated annealing with DEN restraints[13] on the blind tests and on the complete benchmark set of structures related to PDB entries 1XVQ and 1A2B. Using the same template and placement used by Rosetta refinement, initial homology models were built in Modeller[29] (using the same alignment used by Rosetta). DEN refinements were carried out using the refine_lowres.inp script distributed

with CNS version 1.3 as a template. The results of these analyses for the benchmark set of structures are shown in Supplementary Tables 4 and 5, and for the blind tests, as part of Table 1.

Massive-sampling simulated annealing. To test the role that massive sampling around the conformation of the input structure plays in the success of our new methods, we developed an 'extreme simulated annealing protocol', where 1,000 models were produced by simulated annealing refinement, the best of these models is used as the starting point for automated model rebuilding, density modification and refinement with PHENIX, and the resulting model is used as the starting point for a second iteration of the procedure. In this procedure, simulated annealing was carried out in phenix.refine using the flag 'simulated_annealing = True' and the default starting temperature of 5,000 K.

Implementation in Phenix and Rosetta. The procedures described here require considerable computation as up to several thousand Rosetta models are generated for each structure, typically requiring 0.5–1 h per structure of CPU time. We have developed automated procedures in Phenix (phenix.mr_rosetta) that use Rosetta and Phenix modules to carry out and extend many of the methods described here with density modification and density averaging, potentially allowing fewer Rosetta models to be used. Beginning with correctly placed templates (including all copies of each molecule, and placed domains for 13), each of 13 blind test cases in Table 1 can be solved with phenix.mr_rosetta using 20 Rosetta models during each rebuilding cycle, yielding free R values of 0.42 or lower (mean $R_{free} = 0.33$), and requiring from approximately 30 to 130 CPU-hours to complete.

All the methods described in this paper are available in release 3.2 of Rosetta. An application, 'mr_protocols,' is included which was used (together with Phaser and Phenix Autobuild) to generate all the results in this paper. The flags files used for Rosetta are shown below.

Comparative modelling (with target sequence target.fasta, alignment target_template.ali, and template template.pdb) in the context of density:

```
-database $DB
-MR:mode cm
-in:file:extended_pose 1
-in:file:fasta target.fasta
-in:file:alignment target_template.ali
-in:file:template_pdb template.pdb
-loops:frag_sizes 9 3 1
-loops:frag_files aa1xxx_09_05.200_v1_3.gz aa1xxx_03_05.200_v1_3.gz none
-loops:random_order
-loops:random_grow_loops_by 5
-loops:extended
-loops:remodel quick_ccd
-loops:relax relax
-relax:default_repeats 4
-relax:jump_move true
-edensity:mapreso 3.0
-edensity:grid_spacing 1.5
-edensity:mapfile target.map
-edensity:sliding_window_wt 1.0
-edensity:sliding_window 5
-cm:aln_format grishin
-MR:max_gaplength_to_model 10
-nstruct $STRUCTS
```

In cases where Rosetta was used to 'pre-refine' the structure before Phaser, the same command line was used without the -edensity:* flags. Modelling with symmetry used the flags above in addition to the flag '-symmetry_definition symm.def', where symm.def defines the symmetry in the template. Symmetry definition file creation is automated using a script; see the Rosetta documentation for more details.

Additional refinement (both after comparative modelling and after autobuilding in some cases):

```
-database $DB
-MR:mode relax
-in:file:s rosetta_model.pdb
-relax:default_repeats 4
-relax:jump_move true
-edensity:mapreso 3.0
-edensity:grid_spacing 1.5
-edensity:mapfile target.map
-edensity:sliding_window_wt 1.0
-edensity:sliding_window 5
-nstruct 5
```

Comparative modelling against the Patterson function (the experimental Patterson map, target_pat.map, is computed outside Rosetta):

```
-MR:mode cm
-in:file:extended_pose 1
-in:file:fasta target.fasta
-in:file:alignment target_template.ali
-in:file:template_pdb template.pdb
-loops:frag_sizes 9 3 1
-loops:frag_files aa1xxx_09_05.200_v1_3.gz aa1xxx_03_05.200_v1_3.gz none
-loops:random_order
-loops:random_grow_loops_by 5
-loops:extended
-loops:remodel quick_ccd
-loops:relax relax
-relax:default_repeats 2
-relax:jump_move true
-edensity:grid_spacing 1.6
-edensity:mapfile target_pat.map
-edensity:use_spline_interpolation true
-edensity:realign random
-edensity:use_symm_in_pcalc true
-edensity:patterson_lowres_limit 3.5
-edensity:patterson_hires_limit 10.0
-edensity:patterson_minR 5.0
-edensity:patterson_maxR 14.0
-edensity:patterson_B 0.2
-edensity:patterson_cc_wt 0.5
-cm:loop_rebuild_filter 500
-cm:aln_format grishin
-cm:max_loop_rebuild 10
-cm:min_loop_size 4
-MR:max_gaplength_to_model 10
-nstruct $STRUCTS
```

Most of the data used in this paper is available at http://www.phenix-online.org/phenix_data/terwilliger/rosetta_2011/ (additional blind cases will be made available as the structures are deposited).

25. DiMaio, F., Tyka, M. D., Baker, M. L., Chiu, W. & Baker, D. Refinement of protein structures into low-resolution density maps using Rosetta. *J. Mol. Biol.* **392,** 181–190 (2009).
26. André, I., Bradley, P., Wang, C. & Baker, D. Prediction of the structure of symmetrical protein assemblies. *Proc. Natl Acad. Sci. USA* **104,** 17656–17661 (2007).
27. Weis, W. I., Brünger, A. T., Skehel, J. J. & Wiley, D. D. Refinement of the influenza virus hemagglutinin by simulated annealing. *J. Mol. Biol.* **212,** 737–761 (1990).
28. Abe, H., Braun, W., Noguti, T. & Gō, N. Rapid calculation of first and second derivatives of conformational energy with respect to dihedral angles for proteins general recurrent equations. *Comput. Chem.* **8,** 239–247 (1984).
29. Eswar, N. *et al.* Comparative protein structure modeling with MODELLER. *Curr. Protoc. Bioinform.* (Suppl.) **15,** 5.6 doi:10.1002/0471250953.bi0506s15 (2006).

CORRIGENDUM

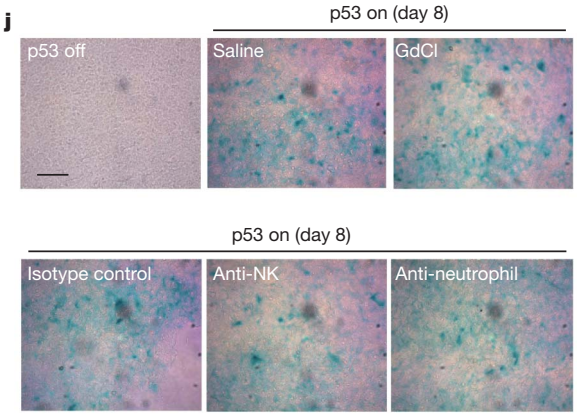
doi:10.1038/nature09909

Senescence and tumour clearance is triggered by p53 restoration in murine liver carcinomas

Wen Xue, Lars Zender, Cornelius Miething, Ross A. Dickins, Eva Hernando, Valery Krizhanovsky, Carlos Cordon-Cardo & Scott W. Lowe

Nature **445,** 656–660 (2007)

In Figure 4j of this Letter, there is a duplicated panel describing the effects of p53 tumour suppressor reactivation in murine liver carcinomas. This figure documented the accumulation of senescence associated β-galactosidase (SA-β-Gal) staining in tumour tissue in response to p53 together with a series of controls. During the final assembly of the manuscript, one panel was inadvertently duplicated such that 'GdCl' treatment and 'isotype control' depict the same image. A corrected figure with the original image for SA-β-gal staining in the isotype control is shown below. The error does not alter the conclusions of the study and the authors apologize for any confusion it may have caused.



CORRIGENDUM

doi:10.1038/nature09991

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium

Nature **467,** 1061–1073 (2010)

In this Article, Yali Xue, of the Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, UK (Analysis group), and Reed A. Cartwright, of the Department of Ecology and Evolutionary Biology, Rice University, Houston, Texas 77251, USA (Analysis group: University of Montreal), were inadvertently omitted from the participant list. Also, the participants David Altshuler, Jonathan Keebler, Paula Kokko-Gonzales and Deborah A. Nickerson were listed incorrectly. In addition, Seungtai C. Yoon should be associated with affiliation 40 (Seaver Autism Center and Department of Psychiatry, Mount Sinai School of Medicine, New York, New York 10029, USA) and not with affiliation 42 (Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine). These have been corrected in the HTML and PDF versions of the manuscript. Supplementary Information section 7.7 has also been corrected.

CORRIGENDUM

doi:10.1038/nature10087

Diphthamide biosynthesis requires an organic radical generated by an iron–sulphur enzyme

Yang Zhang, Xuling Zhu, Andrew T. Torelli, Michael Lee, Boris Dzikovski, Rachel M. Koralewski, Eileen Wang, Jack Freed, Carsten Krebs, Steven E. Ealick & Hening Lin

Nature **465,** 891–896 (2010)

In this Article, the following studies, which led to the identification of diphthamide structure and its biosynthetic genes, should have been cited[1–5]. Reference 1 reports the presence of an unusual amino acid at the ADP-ribosylation site of elongation factor 2 (EF2); ref. 2 reports the properties of the modified residue in EF2 and proposed the name diphthamide; ref. 3 reports the structure of diphthamide; ref. 4 describes work to suggest that the 3-amino-3-carboxypropyl group of diphthamide come from *S*-adenosyl methionine; and ref. 5 reports the identification of yeast mutants that are defective in diphthamide biosynthesis and proposes the biosynthetic pathway.

1. Robinson, E. A., Henriksen, O. & Maxwell, E. S. Elongation factor 2. Amino acid sequence at the site of adenosine diphosphate ribosylation. *J. Biol. Chem.* **249,** 5088–5093 (1974).
2. Van Ness, B. G., Howard, J. B. & Bodley, J. W. ADP-ribosylation of elongation factor by diphtheria toxin: isolation and properties of the novel ribosyl-amino acid and its hydrolysis products. *J. Biol. Chem.* **255,** 10717–10720 (1980).
3. Van Ness, B. G., Howard, J. B. & Bodley, J. W. ADP-ribosylation of elongation factor 2 by diphtheria toxin: NMR spectra and proposed structures of ribosyl-diphthamide and its hydrolysis products. *J. Biol. Chem.* **255,** 10710–10716 (1980).
4. Dunlop, P. C. & Bodley, J. W. Biosynthetic labelling of diphthamide in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **258,** 4754–4758 (1983).
5. Chen, J. Y., Bodley, J. W. & Livingston, D. M. Diphtheria toxin-resistant mutants of *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **5,** 3357–3360 (1985).

CAREERS

EQUALITY Global poll shows gender disparity in physics opportunities **p.547**

US IMMIGRATION Visa process updated for foreign science students **p.547**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

MOODBOARD/CORBIS

Technicians can have duties ranging from washing glassware to doing their own original research.

LAB PERSONNEL

Technically gifted

A good technician can be vital to a successful lab. But how to become one, and how to select the best in a diverse market?

BY HEIDI LEDFORD

When James Damore was looking for a job as a technician, he made it clear from the start that he didn't want to just wash dishes or make reagents. He wanted to do research. He e-mailed a dozen or so professors around the United States to check for openings. "I was direct that I didn't want to do lab chores," he says. "Not many labs had that kind of position."

Damore, fresh from an undergraduate degree in molecular and cellular biology at the University of Illinois at Urbana-Champaign, eventually got a post in Jeff Gore's physics laboratory at the Massachusetts Institute of Technology in Cambridge. He will have spent only about a year there when he heads off to study for a graduate degree in systems biology. Was it worth it for Gore to hire and train a technician with little previous experience and an aversion to mundane lab duties, who would be sticking around for such a short time? "I have two first-author papers with Jeff in press right now," says Damore. "I think I've been worth it."

In academia and industry, lab technicians are often the unsung heroes of successful research programmes. Many, like Damore, are newly out of college and are using the position to boost their prospects of getting into a good graduate programme. Others have more experience and education. Either way, they are in demand, says Alan Edwards, senior director of science at Kelly Scientific Resources, headquartered in Troy, Michigan. "It's a healthy growth market."

The definition of 'technician' varies depending on whether the position is academic or industrial. In academia, the duties can extend from ordering supplies to being a fully fledged researcher. In industry, technicians may also have important roles in manufacturing and quality control.

Accordingly, the educational path can vary. An undergraduate with an eye to pursuing a PhD may want to bolster his or her CV with a short stint as a laboratory technician. Although such students might have little practical experience, they can be highly motivated, and are often attracted to positions that will allow them to gain research experience and a slot on the author's list for any resulting publications.

For others, being a technician is itself a career. Many community colleges offer specialized technician training as part of a bachelor's or master's degree programme, especially in fields such as biomanufacturing. In California alone, there are 32 community-college programmes that train technicians for the biotechnology industry, says Travis Blaschek-Miller, director of communications at BayBio, an association of Northern California life-science companies based in South San Francisco. Local companies fund some training programmes, he notes. "US industry has a high demand for quality research technicians," he says.

The US Bureau of Labor Statistics estimates that the market for scientific technicians will grow by 12% between 2008 and 2018. In Europe, the number of technicians and associated professionals grew by 23% between 2000 and 2009, according to Eurostat in Luxembourg, which keeps statistical records for the European Commission (see 'A varied market'). Demand differs according to the field, however. Edwards says that scientists on temporary visas are making up an increasing proportion of the US technician workforce in some disciplines, as the need for qualified labour outstrips the homegrown supply. In the United States, biology and environmental-science technicians are particularly sought after. But the market for chemical technicians is dropping as companies downsize and find cheaper labour overseas.

"There are many people that could do lab chores. And not many that can do top-flight research."

In fact, many companies are now structuring their businesses to take advantage of distinct talent pools in different regions of the world. China is flush with molecular biologists, for example, so there is increasing interest in ▶

setting up molecular biology labs there — and hiring biology technicians. Meanwhile, Eastern Europe and Latin America, home to burgeoning clinical-trial markets, are training technicians with that in mind.

But financial pressures have affected the qualities that firms are seeking in technicians. Rather than hiring permanent employees, companies worldwide are increasingly interested in temporary workers who can be easily — and cheaply — let go as needs change. "A contingent workforce is the new paradigm," says Edwards. Many companies also want technicians who can hit the ground running, and that means that they value experience. "To wait six months for somebody to get up to speed is less desirable today" than in the past, says Edwards. Many technician-training programmes place an emphasis on internships in academic and industrial laboratories.

But in academia, principal investigators often place more emphasis on motivation than on experience, leaving room for inexperienced candidates such as Damore, who have a clear interest in a research career. Academic technicians will often be paid less than their industry and government counterparts. In the United States, the average annual salary for biological technicians working in research and development across all sectors — academia, government and industry — was US$44,730 in 2009. Damore earns about $36,000 a year — typical for an academic technician fresh out of college.

MIT

"I have two first-author papers in press. I think I've been worth it."
James Damore

So what is the best way to choose a technician? It is a step that some say is among the most important in getting a laboratory up and running, given that it takes time to attract academics to an unproven lab. But getting the right technician can also be a lengthy process. Months before Richard Baxter arrived at Yale University in New Haven, Connecticut, to set up his chemistry lab last year, he had already started advertising for his first technician. "I knew it would take months simply because of the bureaucracy of the hiring process," he says. Still, the time spent is worth it because a technician can get started immediately, helping to set up the lab and getting projects under way. "Several people told me that technicians helped them get the data for their first grants," says Mamta Tahiliani, a biochemist who this year took a post at New York University's Langone Medical Center in New York City.

A new lab-runner can have specific concerns when picking a technician. Baxter worried that while he was off teaching, his recruit would often be alone in the lab, so he asked the referees for each person that he interviewed to describe the applicant's level of independence (see 'How to choose a technician').

A more experienced technician might chafe at taking orders from a young principal investigator, cautions Joanna Chiu, a molecular geneticist at the University of California, Davis. Still, the key factor for many early-career academics is cost. Chiu estimates that it would be almost twice as expensive to hire a technician who has a PhD as one with only an undergraduate degree. Tahiliani agrees, saying that it would be hard to imagine a PhD-holder being satisfied with an academic job that pays only about $35,000, when industry wages are far higher.

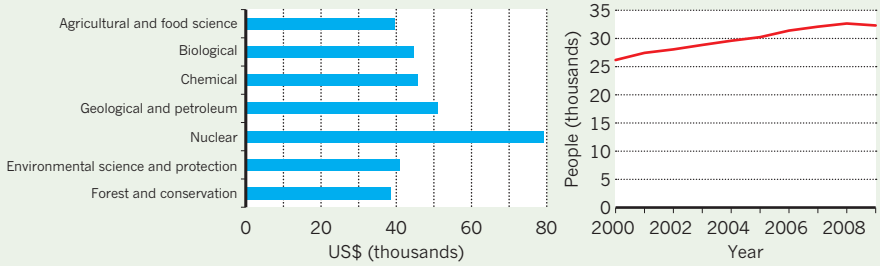
Ultimately, Baxter received applications from both new graduates and experienced technicians with advanced degrees. Although he didn't put a premium on young recruits, he hired a technician just out of university, whose lack of experience hasn't been a problem. "She has rapidly learned to culture cells as reliably as, if not more reliably than, me," he says. "And most important, she's organized and is a friendly face in the lab every day."

Damore's boss, meanwhile, considers his technician's salary to be money well spent. "There are many people that could do lab chores," says Gore. "And not many that can do top-flight research." ■

Heidi Ledford *reports for* Nature *from Cambridge, Massachusetts.*

A VARIED MARKET

Mean salaries for US lab technicians across academia, government and industry vary according to the field in which they work (left). In the European Union, the number of people employed as lab technicians and associated professionals across all sectors and fields has grown since 2000 (right).



SOURCES: LEFT: US BUREAU OF LABOR STATISTICS; RIGHT: EUROSTAT

INTERVIEWING TIPS

How to choose a technician

Questions for applicants

Determining whether a prospective technician is the right fit for the lab can be challenging. Interviewers can test the waters by posing technical questions relevant to the applicant's past work experience. Possible questions include:

- What kind of work would you like to do?
- Would you rather work on one project at a time, or several projects at once?
- Would you be willing to help others with their projects?
- What would you do if something in the lab didn't work?
- Where would you seek help?
- What did and didn't you like about your previous job?
- What kind of supervision did you have?
- With whom did you discuss ideas?
- Did you share equipment with other lab members? How did you prioritize between your work and theirs?
- Talk about a project or situation in your previous job that required initiative. Why did you choose that particular approach?

Questions for referees

Talking to referees can be more revealing than interviewing the candidate, says Joanna Chiu, a molecular geneticist at the University of California, Davis. Mamta Tahiliani, a biochemist at New York University's Langone Medical Center in New York City, agrees. But she cautions that referees tend to be very positive, so it's important to pay close attention to negative comments. A few key questions include:

- Can the applicant work independently?
- What is his or her record-keeping style like?
- Do they have good technical skills?
- Is their work reproducible?
- How do they take criticism?
- Do they understand projects easily?
- Would you rehire them? H.L.

EQUALITY

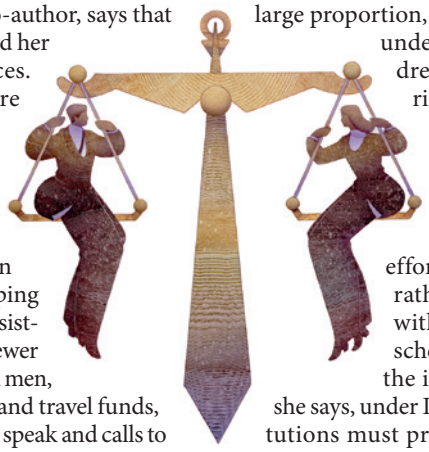
Gender divide in physics spans globe

Report reveals disparity in opportunities and expectations.

BY VIRGINIA GEWIN

An international survey comparing the career experiences of 15,000 physicists from 130 developed and developing nations finds that women around the world experience a tilted playing field. Across the board, the study finds, men have greater access than women to opportunities and resources, and their careers suffer less when they have children.

The survey is the third global poll in a decade to address the experiences of female physicists, but is the first to include men. *Global Survey of Physicists: A Collaborative Effort Illuminates the Situation of Women in Physics* was produced by the American Institute of Physics (AIP) in College Park, Maryland, with funding from the Henry Luce Foundation in New York. Rachel Ivie, assistant director of the AIP's Statistical Research Center and a report co-author, says that the data on men allowed her to compare experiences. "We knew things were unequal, but not this unequal," she says.



J. ENDICOTT/CORBIS

The survey reveals few differences in the degree of gender inequality between developed and developing countries. Women consistently describe getting fewer international offers than men, less access to lab space and travel funds, and fewer invitations to speak and calls to serve on important committees. They also report that having children slows their careers to a greater degree.

Ivie says that two factors contribute to these problems. First, physics remains a male-dominated field, operating through an old boy network. "It's not that senior people actively exclude women; they just don't think of recommending them for key posts or inviting them to speak at conferences," says Ivie.

Elizabeth Freeland, a physics postdoctoral researcher at the University of Illinois at Urbana-Champaign, agrees. "This is an unconscious bias — which makes it harder but not impossible to get past," she says.

The other subtle but sinister factor is that women and men face different cultural expectations. The survey suggests that women are universally considered responsible for childcare and childcare decisions. "The overarching barrier [to women's ascension in the field] is the deeply entrenched perception of both men and women that men are expected to be solely breadwinners, while women are expected to be solely caregivers," says Prajval Shastri, an astrophysicist at the Indian Institute of Astrophysics in Bangalore.

Balancing motherhood and work continues to be the biggest career challenge for women. Carola Meyer, an investigator at the Peter Grünberg Institute in Jülich, Germany, and vice speaker of the German Physical Society's gender-equality working group, says that although institutes and funding bodies provide career breaks for people who wish to have children, such schemes don't necessarily ease the balancing act. Women hold 17% of the 42 positions at her institute — a relatively large proportion, says Meyer. Yet all are under 40 and have no children. Those who want to rise within the scientific community can't consider having children until they are established, she says.

Shastri notes that efforts to provide women, rather than both parents, with childcare or flexible schedules can even add to the inequity. For example, she says, under Indian labour law, institutions must provide childcare if the number of female employees exceeds certain limits. "Laws with no mention of male employees with children effectively imply that women have primary responsibility for children," says Shastri.

Women's perceived roles may also extend to career assignments. Female participation in managerial, editorial or supervisory roles was up to 15% lower than male participation, but in one area women were far more active: advising undergraduates, a 'nurturing' task that typically garners little professional credit.

Meyer is pleased that, for the first time, a study has provided enough statistics to show that the career differences between male and female physicists are universal and deep-rooted. "With such a large survey, the gender differences can't be dismissed," she says. ■

US IMMIGRATION

Student visas extended

More foreign science students are now eligible to stay in the United States for up to 29 months after graduation to gain additional practical training, under a decision announced by the US Department of Homeland Security (DHS) on 12 May. The move adds 50 eligible disciplines, including agriscience, neuroscience and drug design, to the existing list. The DHS — which handles employment visas — says that the change is part of the government's efforts to address shortages of skilled scientists, and the agency expects more applications for extensions as the economy improves. The US visa process has long been criticized for cumbersome delays that have kept many foreign scientists out of the country.

UNITED KINGDOM

Recruits lack skills

Forty-three per cent of UK employers are having trouble recruiting workers with graduate-level skills in science, technology, engineering and maths (STEM), says a report released on 9 May. *Building for Growth: Business Priorities for Education and Skills*, prepared by the Confederation of British Industry (CBI) in London and Education Development International in Coventry, found that employers expect future recruitment problems as numbers of STEM graduates fall. More than one-quarter of science and high-tech employers pay for internships or sponsor higher education to promote STEM to potential recruits, and 60% are increasing investments in training and development. "Employers are taking on a greater role in skills development — offering apprenticeships, training and more links to university programmes," says Simon Nathan, senior policy adviser at the CBI.

EUROPEAN UNION

Boost for networks

A group that backs research collaborations through conferences, exchanges and training has received €30 million (US$43 million) in extra European Commission funding. COST (European Cooperation in Science and Technology) now has a €240-million budget until 2013. Its 250 networks help their 30,000 researcher members across the European Union to get funding from governments and agencies, says Monica Dietl, COST office director. COST also keeps researchers in the region by fostering networking opportunities.

BE SWIFT, MY DARLING

There's no escaping fate.

BY JOHN MORAN

When you wake and start reading this, head fuddled by the cold sleep that crosses the stars, then you're in what the Kree call the saddle (don't worry who the Kree are, my lover, I'll get to that shortly).

For now, just realize that the ring beside you opens an airlock. You'll need to wear the suit because the first two metres are vacuum, but when you jump the gap, and operate the far side by pulling what looks like a purple orchid, you'll find yourself in the alien spacecraft I discovered on my shift three weeks ago (without waking you my darling, I'm sorry).

You'll find this hard, as you enter the glass-smooth tunnel and follow its crimson undulations into the darkness, but you'll also remember this note and tell yourself, as you have all your life, that there's no such thing as fate. You're wrong, my dear, but until the end you won't believe me.

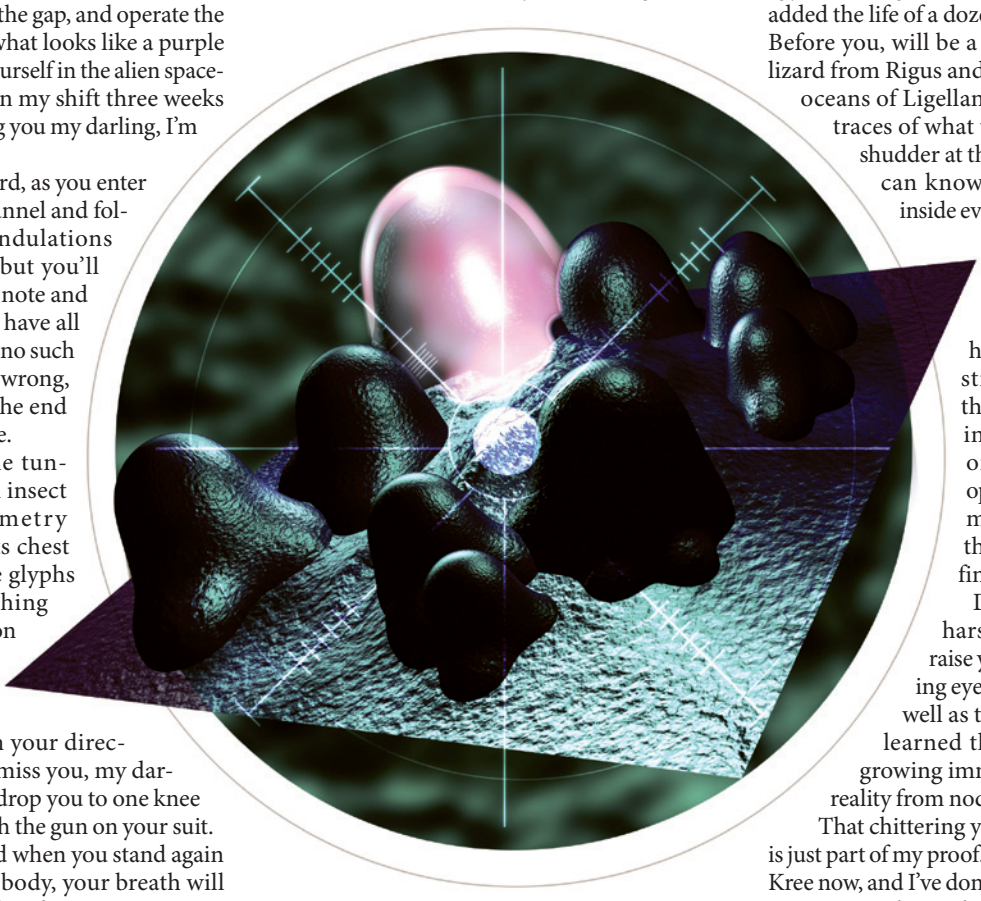
At the end of the tunnel you'll find a tall insect of bilateral symmetry and upright gait, its chest etched with orange glyphs and its screech itching your skin like nails on metal even as one arm turns what appears to be a piece of bamboo in your direction. The shot will miss you, my darling, but shock will drop you to one knee before you reply with the gun on your suit.

This is a Kree, and when you stand again and cross its dying body, your breath will catch as you remember this note.

There are two tunnels. (You think you won't, but you will) take the left-hand one and creep into a red oval that pulsates like a heartbeat. Containers lie here, and some of the Kree survey equipment, but there's no point being wantonly destructive. Just shudder at the way they lie like melted flesh from a Dalí, then make your way through the middle to the heart of the hive.

The Kree are older than us, my darling, and their technology is organic, but through your helmet you won't smell the pheromones keeping the nest viable, or hear the million-pulse beat that shivers from the queen. Instead you'll rush through hanging veils of flesh to a distant glow, amazed that a ship with hundreds of corridors should be so empty, until you remember that I've arranged the timing exactly to give you free access.

You'll panic, then, when you remember what lies in wait, and want to turn back, but you're brave, my darling, so you'll bite your lip and continue even though you'll need the suit's medication to stop you throwing up.



JACEY

At the limit of your explorations you'll think you've entered a cul-de-sac, but you'll remember this note and stretch your fingers into the long, brown grass of the inner ship to discover a hidden passage leading down and know that you've reached the end of your journey.

This last corridor will go by slowly. You'll imagine that each hiss and switch of your intake valve can be heard by the Kree and your skin will feel clammy and then cold as you sweat beyond the capacity of air-conditioning. You'll even begin to tremble when the rippling corridor opens into a damp space filled with black mounds that remind you of eggs.

They aren't, of course, although you won't be able to help your reaction, born of a lifetime of wildlife programmes, till you enter and notice that where the first is leathery, the second is crystalline. The Kree are scavengers, you'll see, and although their technology includes genetic modification, they've added the life of a dozen worlds to the mix. Before you, will be a root from Satir-4, a lizard from Rigus and two clams from the oceans of Ligellan. You'll imagine the traces of what they once were, and shudder at the idea that creatures can know the Kree from the inside even as they are known.

At this point you'll remember your mission and run once more, past half-molten creatures still wailing into the thin air, past fish failing to swim through organic glue, to an open space waiting for more samples. And in the last mound you'll find me.

Don't judge me harshly, darling, as you raise your gun to my pleading eyes, for the Kree give as well as take. From them I've learned that time is organic, growing immovable branches of reality from nodes of choice.

That chittering you'll hear as you aim is just part of my proof. I'm connected to the Kree now, and I've done the mathematics so many times I know that there was no way to bring you in and let you out safely.

Instead, a saddle point. Shoot us both, or let the Kree bind us together in a transformation stranger than you can imagine. I know you expect to put me out of my misery, but here at the end I can only say that you'll have to decide for yourself.

Just begin. Put on the suit and enter the airlock. Be swift, my darling. I'm waiting. ■

NATURE.COM
Follow Futures on Facebook at:
go.nature.com/mtoodm

John Moran *has been a chemical analyst, nuclear physicist and art-shop owner. He currently works as a security consultant for a UK bank.*

LETTER

doi:10.1038/nature10002

Determinants of nucleosome organization in primary human cells

Anton Valouev[1], Steven M. Johnson[2], Scott D. Boyd[1], Cheryl L. Smith[1], Andrew Z. Fire[1,3] & Arend Sidow[1,3]

Nucleosomes are the basic packaging units of chromatin, modulating accessibility of regulatory proteins to DNA and thus influencing eukaryotic gene regulation. Elaborate chromatin remodelling mechanisms have evolved that govern nucleosome organization at promoters, regulatory elements, and other functional regions in the genome[1]. Analyses of chromatin landscape have uncovered a variety of mechanisms, including DNA sequence preferences, that can influence nucleosome positions[2–4]. To identify major determinants of nucleosome organization in the human genome, we used deep sequencing to map nucleosome positions in three primary human cell types and *in vitro*. A majority of the genome showed substantial flexibility of nucleosome positions, whereas a small fraction showed reproducibly positioned nucleosomes. Certain sites that position *in vitro* can anchor the formation of nucleosomal arrays that have cell type-specific spacing *in vivo*. Our results unveil an interplay of sequence-based nucleosome preferences and non-nucleosomal factors in determining nucleosome organization within mammalian cells.

Previous studies in model organisms[3–7] as well as initial analyses in human cells[8] have identified fundamental aspects of nucleosome organization. Here we focus on the dynamic relationships between sequence-based nucleosome preferences and chromatin regulatory function in primary human cells. We mapped tissue-specific and DNA-encoded nucleosome organization across granulocytes and two types of T cells ($CD4^+$ and $CD8^+$) isolated from the blood of a single human donor, by isolating cellular chromatin and treating it with micrococcal nuclease (MNase) followed by deep sequencing of the resulting nucleosome-protected fragments (Methods, Supplementary Fig. 1). To provide sufficient depth for both local and global analyses, we used high-throughput SOLiD technology, generating 584, 342 and 343 million mapped reads for granulocytes, $CD4^+$ and $CD8^+$ T cells, respectively. These are equivalent to 16–28× genome coverage by 147 bp nucleosome footprints (cores; see Methods). The depth of sequence was critical for our subsequent analysis: although shallower coverage can illuminate features of nucleosome positions through statistical analysis (for example, refs 6, 8), any definitive map and thus comparison of static and dynamic positioning requires high sequence coverage throughout the genome.

To provide complementary data on purely sequence-driven nucleosome positioning in the absence of cellular influences, we reconstituted genomic DNA *in vitro* with recombinantly derived histone octamers to produce *in vitro* nucleosomes (Methods, Supplementary Fig. 2), and generated over 669 million mapped reads, representing 32× core coverage of the genome. To identify primary nucleosome positioning sites in DNA, the reconstitution was performed under conditions of DNA excess (see methods). We also generated a control data set of 321 million mapped reads from MNase-digested naked DNA. In the population of granulocytes (our deepest *in vivo* data set), over 99.5% of the mappable genome is engaged by nucleosomes (Methods), and 50 percent of nucleosome-depleted bases occur in regions shorter than 160 bp.

We first focused on global patterns of nucleosome positioning and spacing by calculating fragment distograms and phasograms[6,7,9]. Distograms (histograms of distances between mapped reads' start positions aligning in opposing orientation, Supplementary Fig. 3a) reveal the average core fragment size as a peak if there are many sites in the genome that contain consistently positioned nucleosomes. A positioning signal that is strongly amplified by conditioning the analysis on sites with three or more read starts (reflecting a positioning preference; 3-pile subset), is present not only *in vivo* (Fig. 1a), but also *in vitro* (Fig. 1b), demonstrating that many genomic sites bear intrinsic, sequence-driven, positioning signals. Phasograms (histograms of distances between mapped reads' start positions aligning in the same orientation, Supplementary Fig. 3b) reveal consistent spacing of positioned nucleosomes by exhibiting a wave-like pattern with a period that represents genome-average internucleosome spacing. In granulocytes, the wave peaks are 193 bp apart (Fig. 1c, adjusted $R^2 = 1$, P-value $< 10^{-15}$), which, given a core fragment length of 147 bp, indicates an internucleosome linker length of 46 bp. By contrast, the phasograms of both types of T cells have spacing that is wider by 10 bp (Fig. 1d), equivalent to a 56 bp average linker length. These results are consistent with classical observations of varying nucleosome phases in different cell types[10,11]. Linker length differences have been tied to differences in linker histone gene expression[12,13], which we found to be 2.4 times higher in T cells compared to granulocytes (84 reads per kilobase of mature transcript per million mapped reads (RPKM)[14] vs 35 RPKM). The *in vitro* phasogram (Fig. 1e) reveals no detectable stereotypic spacing of positioned nucleosomes, demonstrating a lack of intrinsic phasing among DNA-encoded nucleosome positioning sites.

Using a positioning stringency metric (Methods; Supplementary Fig. 4) that quantifies the fraction of defined nucleosome positions within a given segment, we calculated the fraction of the genome that is occupied by preferentially positioned nucleosomes at different stringency thresholds. The maximum number of sites at which some positioning preference can be detected statistically is 120 million, covering just over 20% of the genome (Supplementary Fig. 5) at the low stringency of 23%. Thus, the majority of nucleosome positioning preferences is weak, and nucleosomes across the majority of the human genome are not preferentially positioned, either by sequence or by cellular function.

Next we focused on how transcription and chromatin functions affect nucleosome organization regionally. For each cell type, we generated deep RNA-seq data and binned genes into groups according to their expression levels. The average spacing of nucleosomes was greatest within silent genes ($CD4^+$ T cells, 206 bp, Fig. 2a) and decreased by as much as 11 bp as the expression levels went up (t-statistic P-value $= 6.5 \times 10^{-34}$). This suggests that transcription-induced cycles of nucleosome eviction and reoccupation cause denser packing of nucleosomes and slight reduction in nucleosome occupancy (Supplementary Fig. 6). On the basis of this result, we hypothesized that higher-order chromatin organization as implied by specific

[1]Department of Pathology, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, California 94305, USA. [2]Department of Microbiology and Molecular Biology, Brigham Young University, 757 WIDB, Provo, Utah 84602-5253, USA. [3]Department of Genetics, Stanford University School of Medicine, Pasteur Drive, Stanford, California 94305, USA.

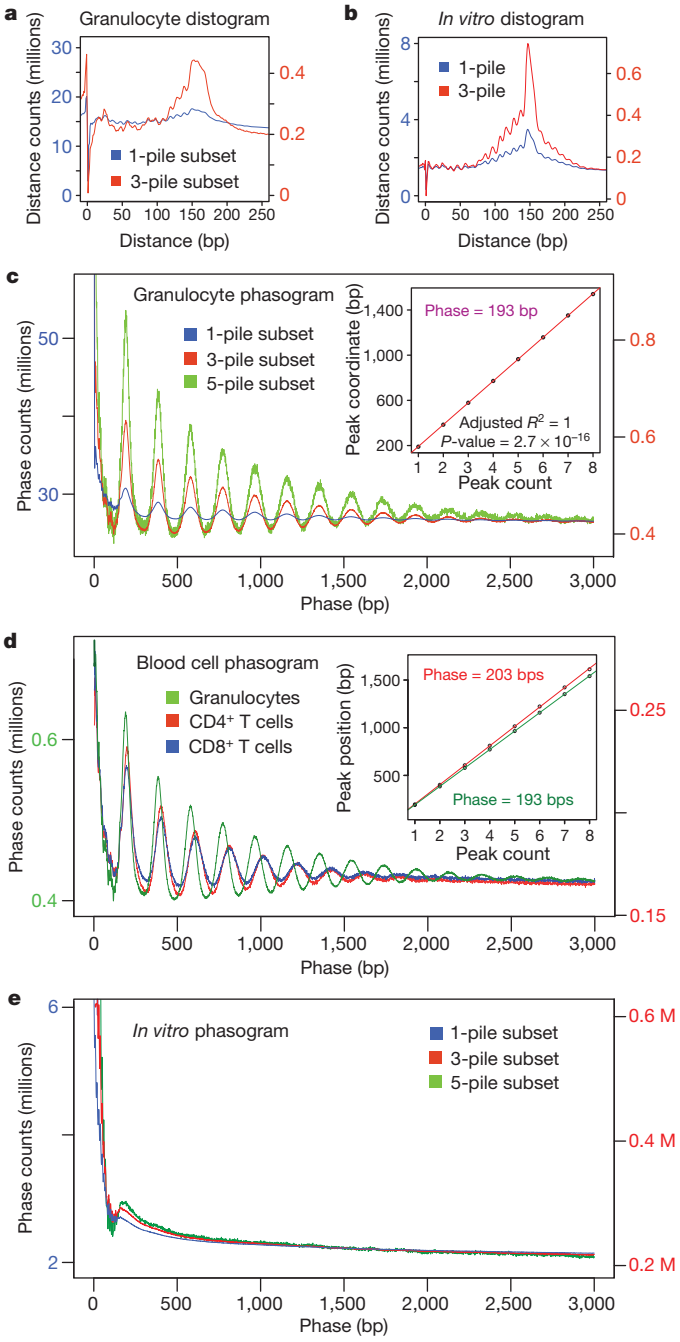


Figure 1 | Global parameters of cell-specific nucleosome phasing and positioning in human. **a**, *In vivo* granulocyte distogram (calculation explained in Supplementary Fig. 3a). *x*-axis represents the range of recorded distances. *y*-axis represents frequencies of observed distances within 1-pile (blue) and 3-pile (red) subsets. 1-pile subset represents the entire data set, 3-pile subset represents a subset of sites containing three or more coincident read starts. **b**, Distogram of the *in vitro* reconstituted nucleosomes showing 1-pile and 3-pile subsets as in (**a**). **c**, *In vivo* granulocyte phasogram (calculation explained in Supplementary Fig. 3b). *x*-axis shows the range of recorded phases. *y*-axis shows frequencies of corresponding phases. Phasograms of 1-pile, 3-pile and 5-pile subsets are plotted. Inset, linear fit to the positions of the phase peaks within 3-pile subsets (slope = 193 bp). **d**, Phasograms of blood cell types. Inset, linear fits in CD4$^+$ T cells (203 bp) and granulocytes (193 bp). **e**, Phasograms of 1-pile, 3-pile and 5-pile subsets in the *in vitro* data.

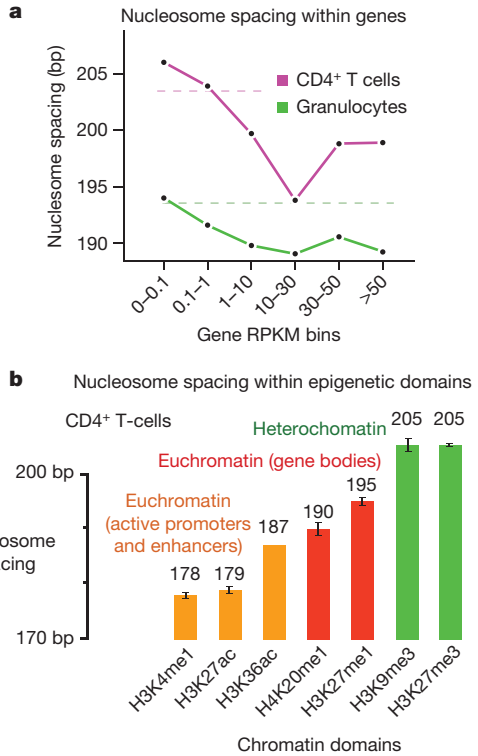


Figure 2 | Transcription and chromatin modification-dependent nucleosome spacing. **a**, Nucleosome spacing as a function of transcriptional activity. *x*- axis represents gene expression values binned according to RPKM values. Internucleosome spacing is plotted along the *y*-axis. Dashed lines represent genome-wide average spacing for each cell type. **b**, Nucleosome spacing within genomic regions marked by specific histone marks in CD4$^+$ T cells. Bar height plots estimated nucleosome spacing for each histone modification. Bar colours differentiate chromatin types (euchromatin vs heterochromatin).

chromatin modifications might be associated with specific spacing patterns. Using previously published ChIP-seq data, we identified regions of enrichment[15] for histone modifications that are found within heterochromatin (H3K27me3, H3K9me3)[16], gene-body euchromatin (H4K20me1, H3K27me1)[16], or euchromatin associated with promoters and enhancers (H3K4me1, H3K27ac, H3K36ac)[17], and estimated spacing of nucleosomes for each of these epigenetic domains. We found that active promoter-associated domains contained the shortest spacing of 178–187 bp, followed by a larger spacing of 190–195 bp within the body of active genes, whereas heterochromatin spacing was largest at 205 bp (Fig. 2b). These results reveal striking heterogeneity in nucleosome organization across the genome that depends on global cellular identity, metabolic state, regional regulatory state, and local gene activity.

To characterize DNA signals responsible for consistent positioning of nucleosomes, we identified 0.3 million sites occupied *in vitro* by nucleosomes at high stringency (>0.5; Methods). The region occupied by the centre of the nucleosome (dyad) exhibits a significant increase in G/C usage (Poisson *P*-value $< 10^{-100}$; Fig. 3a). Flanking regions increase in A/T usage as the positioning strength increases (Fig. 3b). A subset of *in vitro* positioned nucleosomes (stringency > 0.5) which are also strongly positioned *in vivo* (stringency > 0.4) revealed increased A/T usage within the flanks (Fig. 3c) compared to *in vitro*-only positioning sites (Fig. 3a), which underscores the importance of flanking repelling elements for positioning *in vivo*. We term such elements with strong G/C cores and A/T flanks 'container sites' to emphasize the proposed positioning mechanism (Fig. 3d). This positioning signal is different from a 10-bp dinucleotide periodicity observed in populations of nucleosome core segments isolated from a variety of species[18,19] and proposed to contribute to precise positioning and/or rotational setting of DNA on nucleosomes[19] on a fine scale (Supplementary Fig. 7). G/C-rich signals are known to promote nucleosome occupancy[20,21], whereas AA-rich sequences repel nucleosomes[4], and our data demonstrate that precise arrangement of a core-length attractive segment flanked by repelling sequences can produce a strongly positioned nucleosome (Fig. 3d).

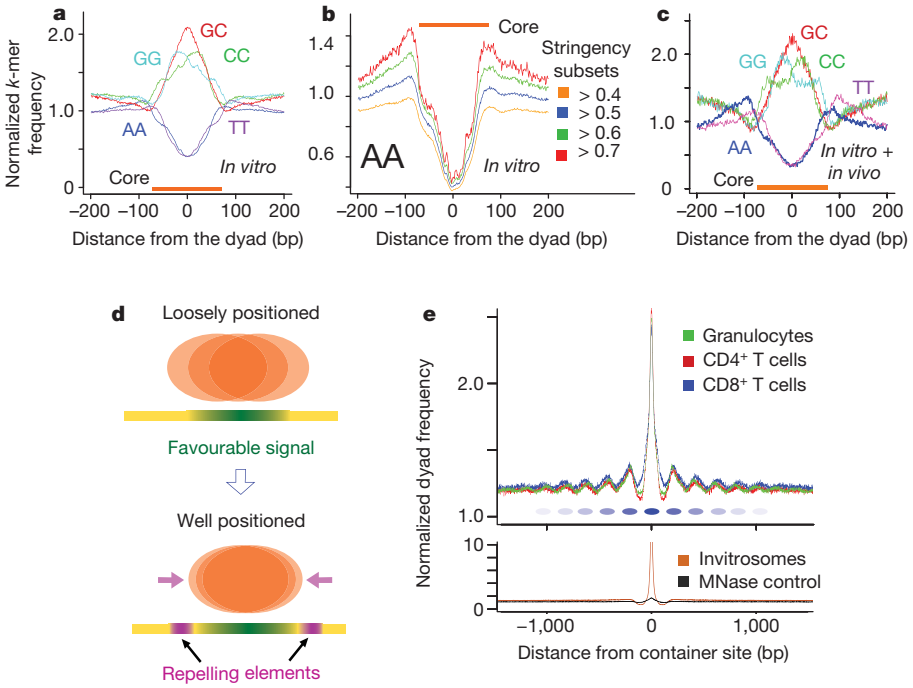


Figure 3 | Sequence signals that drive nucleosome positioning. **a**, Sequence signals within sites containing moderately positioned *in vitro* nucleosomes (stringency > 0.5). Distance from the positioned dyad to a given dinucleotide is plotted along the *x*-axis; *y*-axis represents frequency of a given *k*-mer divided by its genome-wide expectation. The 147-bp footprint of a nucleosome is indicated by an orange band. **b**, Changes in AA dinucleotide usage with increasing positioning stringency. *x* and *y* axes same as in (**a**). Curves of AA usage within the sites of increasingly positioned dyads are shown (stringency cutoffs of 0.4, 0.5, 0.6, 0.7). **c**, Sequence signals within sites containing *in-vitro*-positioned nucleosomes (stringency > 0.5) that also have high *in vivo* stringency (stringency > 0.4). *x* and *y* axes same as in (**a**). **d**, Schematic depiction of the container site positioning mechanism. The C/G-rich core area (green) favours occupancy, but does not precisely position the nucleosome (top). Adding flanking A/T-rich repelling elements (purple, bottom) restricts the position of the nucleosome. **e**, Nucleosome organization around container sites *in vivo* and *in vitro*. *x*-axis represents distances from the dyads to container sites (based on 300,000 container sites). Frequencies of nucleosome dyads around those sites are plotted along the *y*-axis. The upper plot shows distribution of *in vivo* dyads across $CD4^+$ cells, $CD8^+$ cells and granulocytes. The ovals depict hypothetical nucleosome positions across the site with colour intensities reflecting their positioning strength. The lower plot shows distribution of dyads *in vitro* and in MNase control.

Dyad frequencies around container sites (Fig. 3e) show a strong peak of enrichment *in vivo*, confirming that DNA positions nucleosomes *in vivo* over these sites. Additionally, wave-like patterns emanate from these sites *in vivo* (but not *in vitro*), reflecting the nucleation of phased arrays by positioned cellular nucleosomes. Viewing these results in light of the nucleosome barrier model[22], which proposes that nucleosomes are packed into positioned and phased arrays against a chromatin barrier, we conclude that sequence-positioned nucleosome can initiate propagation of adjacent stereotypically positioned nucleosomes. Importantly, wave periods around container sites are shorter in granulocytes than in T cells, allowing tissue-specific variation in linker length (Fig. 1d) to alter placement of nucleosomes over distances of as much as 1 kilobase from an initial container site. Functional consequences of such rearrangements might include global shifts in regulatory properties that could contribute to distinct transcription factor accessibility profiles in different cell types.

The cellular environment can drive nucleosomes to sequences not intrinsically favourable to being occupied, as is evident in a genome-wide comparison of observed nucleosome coverage of all possible tetranucleotides between the granulocyte and the *in vitro* data (Fig. 4a). *In vitro*, nucleosome occupancy is strongly associated with AT/GC content, but this preference is abolished *in vivo*; the exception are C/G rich tetramers that contain CpG dinucleotides, which show a 30% reduction in apparent nucleosome occupancy despite having high core coverage *in vitro*. Consistent with this, CpG islands are fivefold depleted for observed nucleosome coverage *in vivo* (Fig. 4b). No such decrease is observed in the *in vitro* data set.

The decreased nucleosome occupancy of promoters could be due to promoter-related functions of mammalian CpG islands, similar to promoter-associated nucleosome-free regions observed in flies[23] and yeast[5], which do not have CpG islands. We therefore analysed transcription-dependent nucleosome packaging around promoters. As in other organisms[23–27], promoters of active genes have a nucleosome-free region (NFR) of about 150 bp overlapping the transcriptional start site and arrays of well-positioned and phased nucleosomes that radiate from the NFR (Fig. 4c). A notable reduction in apparent nucleosome occupancy extends up to 1 kb into the gene body. We also observed consistent nucleosome coordinates in an independent data set of H3K4me3-bearing nucleosomes[16] (Fig. 4d). Comparison of the nucleosome data (Fig. 4d) with binding patterns of RNA polymerase II[16] (Fig. 4d) around active promoters indicates that phasing of positioned nucleosomes can be explained by packing of nucleosomes against Pol II stalled at the promoter, with Pol II potentially acting as the 'barrier'. The set of inactive promoters, by contrast, exhibits neither a pronounced depletion of nucleosomes, nor a positioning and phasing signal (Fig. 4c). The transition of an inactive promoter to an active one is therefore likely to involve eviction of nucleosomes, coupled with positioning and phasing of nucleosomes neighbouring RNA Pol II (Fig. 4e). These results indicate that CpG-rich segments in mammalian promoters override intrinsic signals of high nucleosome affinity (Supplementary Fig. 8) to become active; this would be in contrast to fly and yeast, where AT-rich promoters may comprise intrinsic sequence signals that are particularly prone to nucleosome eviction[28].

To explore how regulatory factors interact with sequence signals to influence nucleosome organization outside of promoters, we focused on binding sites of the NRSF/REST repressor protein[15] and the insulator protein CTCF. NRSF and CTCF sites are flanked by arrays of positioned nucleosomes (Fig. 4g and Supplementary Fig. 9), consistent with barrier-driven packing previously reported for CTCF[29,30]. Both proteins occupy additional linker space, with NRSF taking up an extra

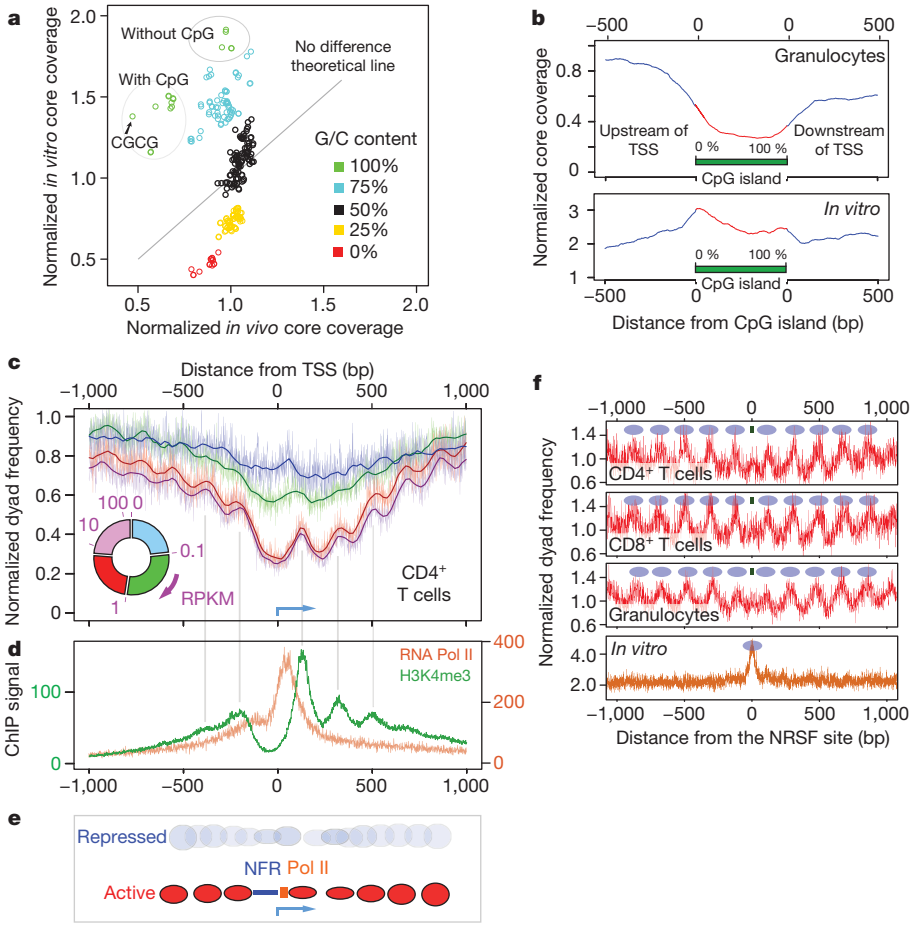


Figure 4 | Influence of gene regulatory function on nucleosome positioning. **a**, Comparison of sequence preferences of nucleosomes *in vivo* and *in vitro*. Normalized nucleosome core coverage *in vivo* (granulocytes) for a given sequence 4-mer is plotted along the *x*-axis. *In vitro* core coverage is plotted along the *y*-axis. Each data point on the plot represents one of the 256 possible 4-mers (coloured according to their G/C content). The diagonal line depicts the positions in the plot for which sequence-based preferences of nucleosomes would be the same *in vivo* and *in vitro*. **b**, Nucleosome core coverage over CpG islands *in vivo* and *in vitro*. *x*-axis represents coordinates within CpG islands (0–100%) and flanking upstream of the transcriptional start sites (TSS) (left) and downstream of the TSS (right). Normalized frequencies of nucleosome cores *in vivo* (upper plot) and *in vitro* (lower plot) are plotted along the *y*-axis. **c**, *In vivo* $CD4^+$ T-cell nucleosome organization around promoters. *x*-axis represents distance from the TSS (blue arrow). Normalized frequencies of nucleosome dyads are plotted along the *y*-axis. Nucleosome arrangements within four gene groups are shown (not expressed 0–0.1 RPKM, low expressed 0.1–1 RPKM, moderately expressed 1–8 RPKM, highly expressed > 8 RPKM). Pie chart depicts distribution of RPKM values across gene groups. **d**, RNA Pol II binding signal within highly expressed genes (orange curve) and H3K4me3-marked nucleosome dyad frequency (green curve) within highly expressed genes (>8 RPKM). Nucleosomes show consistent positions, indicated by grey lines pointing to nucleosome centres. **e**, Schematic depiction of nucleosome organization around promoters of repressed and active genes. Promoters of repressed genes do not have a well-defined nucleosome organization, whereas promoters of active genes have a nucleosome-free region (NFR, blue), RNA Pol II (orange) localized at the NFR boundary, and positioned nucleosomes (red) radiating from the NFR. Height of the ovals represents nucleosome frequency (inferred from **c**). **f**, Nucleosome distribution around the top 1,000 NRSF sites *in vivo* and *in vitro*. Distances from the NRSF binding sites are plotted along the *x*-axis. *y*-axis represents the normalized frequency of nucleosome dyads. Blue ovals depict hypothetical nucleosome positions. NRSF binding site is shown by the green rectangle.

37 bp and CTCF 74 bp. In agreement with sequence-based predictions[21], both CTCF and NRSF sites intrinsically encode high nucleosome occupancy as can be seen from the *in vitro* data (Fig. 4f and Supplementary Fig. 9), but this signal is overridden *in vivo* by occlusion of these sites from associating with nucleosomes. Additionally, phasing of nucleosomes around these regulatory sites is more compact in granulocytes compared to T cells (Supplementary Fig. 9), again exemplifying the importance of cellular parameters for placement of nucleosomes.

Our genome-wide, deep sequence data of nucleosome positions facilitated an initial characterization of the determinants of nucleosome organization in primary human cells. Spacing of nucleosomes differs between cell types and between distinct epigenetic domains in the same cell type, and is influenced by transcriptional activity. We confirm positioning preferences in regulatory elements such as promoters and chromatin regulator binding sites, but find that the majority of the human genome exhibits little if any detectable positioning. The influence of sequence on positioning of nucleosomes *in vivo* is modest but detectable. Despite DNA sequence being a potent driver of nucleosome organization at certain sites, the cellular environment often overrides sequence signals and can drive nucleosomes to occupy intrinsically unfavourable DNA elements or evict nucleosomes from intrinsically favourable sites. We find evidence for the barrier model for nucleosome organization, and that barriers can be nucleosomes (positioned by container sites), RNA polymerase II (stalled at the promoter), or sequence-specific regulatory factors. Our nucleosome maps should be useful for investigating how nucleosome organization affects gene regulation and vice versa, as well as for pinpointing the mechanisms driving regional heterogeneity of nucleosome spacing.

METHODS SUMMARY

Neutrophil granulocytes, $CD4^+$ and $CD8^+$ T cells were isolated from donor blood using Histopaque density gradients and Ig-coupled beads against blood cell surface makers (pan T and $CD4^+$ microbeads, Miltenyi Biotec). Nucleosome cores

were prepared as described previously[7]; cells were snap-frozen and crushed to release chromatin, followed by micrococcal nuclease treatment. *In vitro* nucleosomes were prepared by combining human genomic DNA with recombinantly-derived histone octamers at an average ratio of 1 octamer per 850 bp. Unbound DNA was then digested using micrococcal nuclease. After digestion, reactions were stopped with EDTA, samples were treated with proteinase K, and nucleosome-bound DNA was extracted with phenol-chloroform and precipitated with ethanol (Supplementary methods). Purified DNA was size-selected (120–180 bp) on agarose to obtain mononucleosome cores, followed by sequencing library construction. RNA was isolated by homogenizing purified cells in TRIzol, poly-A RNA was purified using a Qiagen Oligotex kit and RNA-seq libraries were constructed using a SOLiD Whole Transcriptome Analysis kit. All sequence data was obtained using the SOLiD 35 bp protocol and aligned using the SOLiD pipeline against the human hg18 reference genome. Downstream analyses were all conducted using custom scripts (Methods).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 11 August 2010; accepted 18 March 2011.
Published online 22 May 2011.

1. Mellor, J. The dynamics of chromatin remodeling at promoters. *Mol. Cell* **19,** 147–157 (2005).
2. Radman-Livaja, M. & Rando, O. J. Nucleosome positioning: how is it established, and why does it matter? *Dev. Biol.* **339,** 258–266 (2010).
3. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458,** 362–366 (2009).
4. Berstein, B. E., Liu, C. L., Humphrey, E. L., Perlstein, E. O. & Schreiber, S. L. Global nucleosome occupancy in yeast. *Genome Biol.* **5,** R62 (2004).
5. Yuan, G.-C. *et al.* Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309,** 626–630 (2005).
6. Johnson, S. M., Tan, F. J., McCullough, H. L., Riordan, D. P. & Fire, A. Z. Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. *Genome Res.* **16,** 1505–1516 (2006).
7. Valouev, A. *et al.* A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18,** 1051–1063 (2008).
8. Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132,** 887–898 (2008).
9. Trifonov, E. N. & Sussman, J. L. The pitch of chromatin DNA is reflected in it its nucleotide sequence. *Proc. Natl Acad. Sci. USA* **77,** 3816–3820 (1980).
10. Kornberg, R. D. Structure of chromatin. *Ann. Rev. Biochem.* **46,** 931–954 (1977).
11. Widom, J. A relationship between the helical twist of DNA and the ordered positioning of nucleosomes in all eukaryotic cells. *Proc. Natl Acad. Sci. USA* **89,** 1095–1099 (1992).
12. Schlegel, R. A., Haye, K. R., Litwack, A. H. & Phelps, B. M. Nucleosome repeat lengths in the definitive erythroid series of the adult chicken. *Biochim. Biophys. Acta* **606,** 316–330 (1980).
13. Fan, Y. *et al.* Histone H1 depletion in mammals alters global chromatin structure but causes specific changes in gene regulation. *Cell* **29,** 1199–1212 (2005).
14. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (2008).
15. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods* **5,** 829–834 (2008).
16. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837 (2007).
17. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genet.* **40,** 897–903 (2008).
18. Satchwell, S. C., Drew, H. R. & Travers, A. A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191,** 659–675 (1986).
19. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442,** 772–778 (2006).
20. Hughes, A. & Rando, O. J. Chromatin 'programming' by sequence - is there more to the nucleosome code than %GC? *J. Biol.* **8,** 96 (2009).
21. Tillo, D. *et al.* High nucleosome occupancy is encoded at human regulatory sequences. *PLoS ONE* **5,** e9129 (2010).
22. Mavrich, T. N. *et al.* A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.* **18,** 1073–1083 (2008).
23. Mavrich, T. N. *et al.* Nucleosome organization in the *Drosophila* genome. *Nature* **453,** 358–362 (2008).
24. Lee, W. *et al.* A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genet.* **39,** 1235–1244 (2007).
25. Gu, S. G. & Fire, A. Partitioning the *C. elegans* genome by nucleosome modification, occupancy, and positioning. *Chromosoma* **119,** 73–87 (2010).
26. Sasaki, S. *et al.* Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. *Science* **323,** 401–404 (2009).
27. Zhang, Y. *et al.* Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions *in vivo*. *Nature Struct. Mol. Biol.* **16,** 847–852 (2009).
28. Field, Y. *et al.* Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. *Nature Genet.* **41,** 438–445 (2009).
29. Chuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.* **19,** 24–32 (2009).
30. Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet.* **4,** e1000138 (2008).
31. Albert, I. *et al.* Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* **446,** 572–576 (2007).
32. Wellinger, R. E. & Thoma, F. Nucleosome structure and positioning modulate nucleotide excision repair in the non-transcribed strand of an active gene. *EMBO J.* **16,** 5046–5056 (1997).
33. Sha, K. *et al.* Distributed probing of chromatin structure *in vivo* reveals pervasive chromatin accessibility for expressed and non-expressed genes during tissue differentiation in *C. elegans*. *BMC Genomics* **11,** 465 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the Stanford Genetics/Pathology Sequencing Initiative. We thank G. Narlikar for help with *in vitro* experiments, Life Technologies, especially J. Briggs, for help with generating sequencing data, P. Lacroute for help with sequence alignment, S. Galli for valuable discussions, L. Gracey for critical reading of the manuscript, and members of the Sidow and Fire labs for valuable feedback and discussions. Work in the Fire lab was partially supported by NIGMS (R01GM37706). A.V. was partially supported by an ENCODE subcontract to A.S. (NHGRI U01HG004695). S.M.J. was partially supported by the Stanford Genome Training program (NHGRI T32HG00044).

Author Contributions A.V., S.M.J., A.S. and A.Z.F. designed the experiments. S.M.J., A.V., C.L.S. and S.D.B. performed the experiments. A.V. designed and carried out analyses with input from A.S., A.Z.F. and S.M.J.; A.V., A.S. and A.Z.F. wrote the manuscript.

Author Information All sequence data were submitted to Sequence Read Archive (accession number GSE25133). Sites containing strongly positioned *in vitro* nucleosomes are available as a supplementary data file. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to A.S. (arend@stanford.edu) or A.Z.F. (afire@stanford.edu).

METHODS

Cell purification. Blood samples were obtained from the Stanford Blood Center. Samples were screened for any medical history of malignancy or signs of infectious disease, and tested for serologic evidence of viral infections to ensure that samples came from healthy donors. The Stanford Blood Center procedures used for the cells in this study are the same as those used for transfusion of patients and are routinely inspected by the FDA, the American Association of Blood Banks, and the College of American Pathologists. The blood for the experiments was processed immediately upon donation to avoid any change in quality as a result of sample storage.

Buffy coat (36 ml) from a blood donor was diluted in PBS to a total volume of 200 ml. The cells were layered on a Histopaque gradient with densities 1.119 and 1.077 g ml^{-1} according to manufacturer's instructions (Sigma HISTOPAQUE-1119 and 1077) and separated by centrifugation to yield granulocytes and mononuclear fractions. T cells were isolated from mononuclear cells using a Pan T isolation Kit (Miltenyi Biotec), followed by separation into CD4$^+$ and CD8$^+$ fractions using CD4$^+$ microbeads (Miltenyi Biotec).

Isolation of mononucleosome core DNA fragments from human cells. To isolate mononucleosome core DNA from human cells, neutrophil granulocytes, CD4$^+$ lymphocytes and CD8$^+$ lymphocytes were flash-frozen in liquid nitrogen in 0.34 M sucrose Buffer A and ground, digested on different days, and isolated as described in ref 7. By carrying out an MNase digestion in a short time frame (12 min at 16 °C) following grinding of the samples, we minimize the potential for nucleosome mobility. To maximize uniformity of representation, we use an extraction protocol after MNase digestion that does not rely on solubility of the individual core particles; this resulted in recovery of the bulk of input DNA as a mono-nucleosome band (Supplementary Fig. 10), limiting the degree to which the protocol might select for specific (for example, accessible) chromosomal regions. The mean nucleosome core length obtained for analysis (153 nucleotides) indicates an average overhang of 3 nucleotides on each side of individual cores (147 bp + 2 × 3 bp = 153 bp). Subsequent analyses assign nucleosome positions accounting for this mean overhang and making use of the ability to define location based on interpolation between values calculated from plus-oriented and minus-oriented reads (see below).

Preparation of *in vitro* nucleosomes. Naked genomic DNA isolated from neutrophil granulocytes from our *in vivo* studies was sheared by sonication using a Covaris sonicator and separated on a 1% UltraPure Agarose (Invitrogen) gel run at 100 V for 1 h. A smear of fragments with lengths from 850–2,000 bp (the bulk of the sheared DNA) was isolated and extracted from the gel using the QIAquick Gel Extraction Kit (Qiagen). DNA fragment lengths several-fold larger than nucleosome cores were chosen for this analysis to minimize any end-effects that could have contributed an end-based signal at shorter fragment sizes. Lack of end-preference in the reconstitutions was then confirmed under the conditions of these assays using a series of defined restriction fragments as templates for assembly (S.M.J. and A.F., results not shown).

The ends of the sheared DNA fragments were repaired as described below and then were assembled with recombinant *Xenopus* histones into nucleosomes as described previously[34] at a 1.1:1 molar ratio of DNA to histone octamer such that on average one nucleosome would occupy 850 bp of DNA. Specifically, 4.9 μg of DNA and 0.80 μg of octamer were reconstituted in a total volume of 200 μl.

The ref. 34 conditions (in which DNA was not limiting) were used for our analysis in order to focus specifically on primary sequence effects on nucleosome position. We note that two recent studies in yeast use somewhat different conditions, with a higher ratio of nucleosomes to DNA[3,27]. Assays at high nucleosome:DNA ratio provide a composite readout reflecting both (1) primary preferences of nucleosomes (caused by sequence signals within the nucleosome-bound DNA) and (2) secondary effects due to steric hindrance as a result of dense packing of nucleosomes. Although such data are certainly valuable in modelling chromosome dynamics, the goals of our study (definition of individual sequence elements that can initiate positioning) were best served with the lower nucleosome:DNA assay conditions[34].

Isolation of *in vitro* nucleosome core DNA fragments. *In vitro* nucleosome core DNAs were isolated by diluting 70 μl of the reconstituted *in vitro* nucleosome into a total volume of 200 μl containing 5 mM $MgCl_2$, 5 mM $CaCl_2$, 70 mM KCl and 10 mM Hepes at pH 7.9 (final concentrations) and digesting with 20 units of micrococcal nuclease (Roche) resuspended at 1 U μl^{-1} for 15 min at room temperature. The digestion was stopped by adding an equal volume of 3% SDS, 100 mM EDTA and 50 mM Tris. Octamer proteins were removed by treating with one-tenth volume proteinase K (20 mg ml^{-1} in TE at pH 7.4) for 30 min at 50 °C followed by phenol/chloroform and chloroform extractions and ethanol precipitation. This procedure was repeated twice to process the entire *in vitro* sample, and then *in vitro* DNA cores were isolated on a 2% UltraPure Agarose (Invitrogen) gel run at 100 V for 1 h followed by DNA extraction from the gel using a QIAquick Gel Extraction Kit (Qiagen) following the standard protocol with the exception of allowing the isolated gel sample to incubate in Buffer QG at room temperature until dissolved.

Genomic MNase digest control library preparation. For control libraries, genomic DNA (20 μg) from human neutrophil granulocytes in 0.34 M sucrose Buffer A with 1× BSA (New England Biolabs) and 1 mM $CaCl_2$ was digested with 200 units of micrococcal nuclease (Roche) (0.4 U μl^{-1} final concentration) in a total volume of 500 μl for 10 min at 23 °C. The digestion was stopped by addition of 10 μl 0.5 M EDTA, followed by ethanol precipitation. The digested DNA was run on a 2.5% agarose gel and the smear of DNA fragments from 135–225 bp was excised from the gel and purified using a QIAquick Gel Extraction Kit (Qiagen) as noted above.

End repair, linker ligation and library amplification. The ends of isolated mononucleosome core DNAs (granulocytes, CD4$^+$ lymphocytes and CD8$^+$ lymphocytes), *in vitro* core DNAs and genomic control DNAs were processed by treating 0.3–0.5 μg of the DNA samples with T4 polynucleotide kinase (New England Biolabs) at 37 °C for 2.5 h followed by ethanol precipitation and subsequent treatment with T4 DNA polymerase (New England Biolabs) in the presence of dNTPs for 15 min at 12 °C. After purification using either a QIAquick Gel Extraction Kit as described above or a QIAquick PCR Purification Kit (Qiagen), linkering of previously annealed duplexes AF-SJ-47 (5′-OH-CCACTACGCCT CCGCTTTCCTCTCTATGGGCAGTCGGTGAT-3′)/AF-SJ-48 (5′-P-ATCAC CGACTGCCCATAGAGAGGAAAGCGGAGGCGTAGTGGTT-3′) and AF-SJ-49 (5′-OH-CTGCCCCGGGTTCCTCATTCTCT-3′)/AF-SJ-50 (5′-P-AGAG AATGAGGAACCCGGGGCAGTT-3′) to the samples was accomplished with T4 DNA ligase during a 6.5-h room-temperature incubation. The ligation reactions were separated on a 2% agarose gel, and the relevant band isolated as described above. Amplification of the linkered libraries was accomplished with 8 (granulocyte mononucleosome library), 10 (CD4$^+$ lymphocytes, CD8$^+$ lymphocytes and genomic control libraries) or 12 (*in vitro* library) cycles of polymerase chair reaction (PCR) using primers AF-SJ-47 (SOLiD P1 primer) and AF-SJ-49 (SOLiD P2 primer) with subsequent separation and purification using a 2% agarose gel and the QIAquick Gel Extraction Kit as described above. The number of cycles used in the PCR amplification were monitored and selected as described in ref. 25.

RNA-seq library preparation. Cells were homogenized in TRIzol using an 18G needle, followed by total RNA extraction using phenol-chloroform-isoamyl alcohol. Poly-A RNA was isolated from total RNA using a Qiagen Oligotex kit according to the manufacturer's instructions. The RNA-seq SOLiD sequencing library was built from 100 ng of poly-A RNA according to the manufacturer's instructions (SOLiD whole transcriptome analysis kit).

DNA sequencing and mapping. Both nucleosome fragment and RNA-seq libraries were sequenced using the SOLiD DNA sequencing platform to produce 35 bp reads. All sequence data was mapped using SOLiD software pipeline against the human hg18 assembly using the first 25 bp from each read. This was done to maximize the number of the reference-mapped reads, as the higher error rate in read positions 26–35 of that version of the SOLiD chemistry prevented a substantial fraction of reads from mapping to the genome. For the genome-wide analysis we retained only unambiguously mapped reads.

Genome coverage by nucleosome cores was calculated as: core coverage = (number of mapped reads) × (147)/(genome size)

mRNA sequencing and data analysis. RNA-seq libraries were sequenced on the SOLiD platform to produce 35 bp reads and then the first 25 bp of each read were mapped to hg18 using the SOLiD mapping pipeline which resulted between 77 and 99 million mapped reads for each cell type. RPKM values were calculated as in ref. 14, with a modification that adjusted for transcript length, which was calculated according to the formula $L' = L - 50 \times (E - 1)$, where L is the actual transcript length, and E is the number of exons in the gene. This modification is needed because of the lack of mappings across splice junctions.

Mathematical notations. Start counts: $S_{+/-}(j)$ represent counts of 5′ coordinates of reads that map in + or – orientation at the j-th position of the reference strands. For example, if read maps to the interval $[x,y)$ on the + strand, then its 5′ coordinate is x, if it maps to – strand, then it's $y - 1$.

Indicator functions: I(condition) = 1 if condition is satisfied, 0 otherwise.

Nucleosome positioning stringency metric: nucleosome positioning stringency metric quantifies the fraction of nucleosomes covering a given position that are 'well positioned'. The stringency at position i of the genome is calculated according to the formula:

$$S(i, w=30) = \frac{D(i, w=30)}{\sum_{i-150}^{i+150} \frac{1.09}{w} D(j, w=30)},$$

where $D(i,w)$ is a kernel-smoothed dyad count calculated according to the formula:

$$D(i,w) = \sum_{j=0}^{L} K(i-j,w) \cdot d(j),$$

where L is the size of a given chromosome, and $K(u,w)$ is a smoothing kernel function of the form:

$$K(u,w)=(1-(u/w)^2)^3 I\{|u|<w\},$$

and

$$\int_{-1}^{1}(1-u^2)^3 du = 1/1.09,$$

and $d(j)$ represents the number of dyads that occurs at the position j:

$$d(j)=s_+(j-l/2)+s_-(j+l/2).$$

Here l is the average library size ($l=153$ for *in vivo* data sets, 147 for *in vitro* data set). The core size is inferred from the 3-pile distogram peak position in the range of 100–200 bp.

The numerator of the stringency formula represents a kernel-smoothed count of nucleosome centres (dyads) at position i in the genome, whereas the denominator represents the count of nucleosome centres that infringe on the nucleosome centred at that position, which is inferred by integration of the dyad density estimate over an area of nucleosome infringement. The stringency is constructed in such a way that it would achieve a maximum of 1 if all nucleosomes were perfectly centred at that position (Supplementary Fig. 4). If two alternative, mutually exclusive, equally frequent nucleosome positions are observed in the data, then the stringency would be 0.5 or 50% for each alternative site (illustrated in Supplementary Fig. 4).

Application of the Kernel Density Estimation allowed obtaining smooth estimates of the stringency, which was useful for detection of nucleosome centres and robustly estimating the degree of positioning. We experimented with other smooth kernels and obtained highly consistent results. In principle, the kernel choice should not affect the results substantially as long as there is sufficient nucleosome core coverage (which follows from the convergence property of Kernel Density Estimation).

The kernel bandwidth w is an important parameter of the stringency formula and provides a means to control the smoothness of the stringency profile. Larger values of w provide higher smoothing but result in less accurate estimates of positioning centres, which is acceptable in cases of low core coverage. On the other hand, lower values of w result in less smoothing but more accurate estimation of the positioning centres, which is desirable in cases when nucleosome core coverage is high. We decided to use $w=30$ in our calculation as it provided a sufficient amount smoothing across all of our data sets without sacrificing the sharpness of the positioning estimate.

Nucleosome positioning stringency was used for calculation of the fraction of the genome containing preferentially positioned nucleosomes (Supplementary Fig. 5). Positioned nucleosomes used in the container site analysis (Fig. 3a–c) were identified with the positioning stringency metric (as shown) and additional filters on nucleosome occupancy (*in vitro* occupancy >30) to improve the statistical confidence of the positioning estimates.

Nucleosome dyad coordinates. Nucleosome dyads were inferred from 5′ coordinates of reads by shifting them by half the average nucleosome core size towards the 3′ end. The average nucleosome cores size was estimated by a maximum value of the 3-pile distogram in a size range of 100–200 bp.

Rotational positioning analysis. We examined oligonucleotide preferences of rotational positioning of nucleosomes, which is associated with 10-bp patterning of short k-mers within nucleosome cores[18,31]. Plotting the frequencies of dyads around specific oligomers within the genome showed that the strongest patterning was exhibited by C-polymers (CC,CCC) with an exact helical period of 10.15 bp (Supplementary Fig. 7a, P-value $<2\times10^{-16}$), indicating that they are important for rotational positioning. *In vivo*, such rotational preferences are much less pronounced (Supplementary Fig. 7b), indicating that cellular factors or conditions often override the sequence-encoded rotational settings.

Characterization MNase cleavage patterns. MNase is known to have sequence preferences that can affect both individual and bulk analyses of chromatin structure. Previous studies comparing MNase with alternative probes in model systems, both at specific loci (for example, ref. 32) and genome wide (for example, ref. 33), support the correspondence between the patterns of nucleosomes inferred from MNase digestion of chromatin and the *in vivo* chromatin landscapes. Nonetheless, it remained important to characterize the patterns of MNase activity in our data.

We investigated the extent of cleavage bias by MNase by examining sequence preferences within the cleavage sites, which correspond to 5′ end read positions in our data (Supplementary Figure 7a–e). Consistent with previous observations, MNase exhibits a pronounced but imperfect tendency to cleave at A or T nucleotides in naked DNA (Supplementary Fig. 11a). This same bias is detectable but, importantly, weaker when nucleosomes occupy the DNA, both *in vivo* and *in vitro* (1-pile subsets, top row b–e). Sites of more frequent cleavage (3-pile and 5-pile subsets, middle and bottom rows) revealed preferences that were virtually indistinguishable from the single-site preference.

The fact that the cleavage bias does not extend beyond 1–2 base pairs suggests that our analyses of nucleosome positioning preferences, which have substantially less than single-base resolution, should be robust to biases introduced by the MNase digestion. A case in point is the above-discussed rotational positioning analysis, whose resolution is on the order of 10 bp and which involves oligonucleotides that do not resemble the MNase cleavage site (Supplementary Fig. 7a).

To investigate whether the sequence-driven nucleosome positioning element identified by the *in vitro* reconstitution experiment (Fig. 3) was a result of particularly pronounced MNase digestion bias within specific sites, we examined nucleotide preferences of nucleosome fragments overlapping sites of medium (>0.5) and high (>0.7) positioning stringency (Supplementary Fig. 11f, g). Preferences within these sites are identical to genome-wide preferences, ruling out the possibility that their positioning is an artefact of MNase digestion. In addition, we observe wave-like patterns *in vivo* around these sites (Fig. 3e) consistent with existence of a chromatin barrier in the form of a well-positioned nucleosome.

The lack of systematic differences in cleavage bias in our experimental data sets, in conjunction with the fact that naked DNA is affected most by the cleavage bias, suggests that our conclusions are robust to the use of MNase.

Analyses of independent data sets. We conducted additional analyses on independent data not generated by us to address any lingering concerns about biases or reproducibility. First, we sought to confirm independently that MNase cuts the linker DNA separating nucleosomes. In our data, CTCF sites (Supplementary Fig. 9) are surrounded by arrays of highly positioned and phased nucleosomes extending at least 1 kb in each direction. We investigated the frequency of cleavage by DNase I, a nuclease with preferences different from those of MNase, around CTCF sites within lymphoblastoid cell lines, using publicly available data from the ENCODE project. In agreement with our MNase results, we observed strongly phased peaks in the DNase I ENCODE data that align with linker DNA sites in our nucleosome data (Supplementary Fig. 12).

The estimates of spacing between nucleosomes as depicted in Fig. 1d are consistent between the two types of T cells we analysed. To ask whether these estimates were also reproducible by a different approach, we turned to a published data set that was generated for a different purpose, and by different means. Ref. 8 compared nucleosome distribution between resting and activated $CD4^+$ T cells using MNase treatment of the cellular chromatin. We analysed spacing of nucleosomes in their data and obtained a highly concordant estimate of 202 and 203 bp (Supplementary Fig. 13) which is in agreement with the 203 bp spacing we see in our data (Fig. 1d).

34. Luger, K., Rechsteiner, T. J. & Richmond, T. J. Preparation of nucleosome core particle from recombinant histones. *Methods Enzymol.* **304,** 3–19 (1999).

LETTER

doi:10.1038/nature10015

Tunable pK_a values and the basis of opposite charge selectivities in nicotinic-type receptors

Gisela D. Cymes[1] & Claudio Grosman[1]

Among ion channels, only the nicotinic-receptor superfamily has evolved to generate both cation- and anion-selective members. Although other, structurally unrelated, neurotransmitter-gated cation channels exist, no other type of neurotransmitter-gated anion channel, and thus no other source of fast synaptic inhibitory signals, has been described so far. In addition to the seemingly straightforward electrostatic effect of the presence (in the cation-selective members) or absence (in the anion-selective ones) of a ring of pore-facing carboxylates, mutational studies have identified other features of the amino-acid sequence near the intracellular end of the pore-lining transmembrane segments (M2) that are also required to achieve the high charge selectivity shown by native channels[1–10]. However, the mechanism underlying this more subtle effect has remained elusive[11] and a subject of speculation. Here we show, using single-channel electrophysiological recordings to estimate the protonation state of native ionizable side chains, that anion-selective-type sequences favour whereas cation-selective-type sequences prevent the protonation of the conserved, buried basic residues at the intracellular entrance of the pore (the M2 0′ position). We conclude that the previously unrecognized tunable charge state of the 0′ ring of buried basic side chains is an essential feature of these channels' versatile charge-selectivity filter.

The amino-acid differences that underlie the opposite charge selectivities of the members of the nicotinic-receptor superfamily have been known for several years now[1–10] (see Supplementary Fig. 1 and Supplementary Text for a brief introduction to this group of ion channels). Reversal-potential measurements have revealed that the cation-selective members of the superfamily become anion selective upon both insertion of a proline into the loop that connects transmembrane segments M1 and M2 (between positions −2′ and −1′), and mutation of the pore-lining glutamate at position −1′ to alanine, two changes that bring the sequence of the cation channels closer to that of their anion-selective counterparts (Fig. 1a). Similarly, the reciprocal changes (that is, deletion of the −2′ proline and mutation of the −1′ alanine to glutamate) engineered on (natively) anion-selective members of the superfamily have been shown to confer high selectivity for cations. Whereas the effect of the presence or absence of pore-exposed carboxylates may seem unsurprising, the basis for the effect of the insertion or deletion of a proline from the M1–M2 loop on charge selectivity (Fig. 1b) is much more subtle and has remained, largely, a mystery[11].

As a first step towards understanding the effect of the proline insertion, we engineered a proline between positions −2′ and −1′ of the α1, β1 and δ subunits of the (cation-selective) $(\alpha1)_2\beta1\delta\varepsilon$ acetylcholine receptor (muscle AChR), one subunit at a time. Because the ε subunit already has an extra residue at this position (a glycine; Fig. 1a), a proline was introduced in this subunit by a residue-to-residue mutation rather than by insertion. In the presence of a typical, divalent cation-containing solution in the pipette of cell-attached patches ($[Ca^{2+}] = 1.8$ mM; $[Mg^{2+}] = 1.7$ mM; solution 1 in Supplementary Table 1), single-channel recordings from the proline-insertion mutant in, for example, the δ subunit show an unusually noisy open-channel

a

		Intracellular end of M1			M1–M2 loop								Intracellular end of M2		
Subunit	Origin	−10′	−9′	−8′	−7′	−6′	−5′	−4′	−3′	−2′	−1′	0′	1′	2′	3′
AChRα1	Mouse	F	Y	L	P	T	D	S	-	G	E	K	M	T	L
AChRα1	Zebrafish	F	Y	L	P	T	D	S	-	G	E	K	M	T	L
AChRα1	*A. californica*	F	Y	L	P	S	D	S	-	G	E	K	I	T	L
AChRα1	Fruit fly	F	Y	L	P	S	D	S	-	G	E	K	I	S	L
AChRα4	Mouse	F	Y	L	P	S	E	C	-	G	E	K	V	T	L
AChRα7	Mouse	F	L	L	P	A	D	S	-	G	E	K	I	S	L
AChRβ1	Mouse	F	Y	L	P	P	D	A	-	G	E	K	M	G	L
AChRδ	Mouse	F	Y	L	P	G	D	C	-	G	E	K	T	S	V
AChRε	Mouse	Y	F	L	P	A	Q	A	G	G	Q	K	C	T	V
5-HT_{3A}R	Mouse	F	C	L	P	P	D	S	-	G	E	R	V	S	F
GlyLRα1	Mouse	F	W	I	N	M	D	A	A	P	A	R	V	G	L
$GABA_A$Rα1	Mouse	F	W	L	N	R	E	S	V	P	A	R	T	V	F
GluClα1	*C. elegans*	F	W	F	D	R	T	A	I	P	A	R	V	T	L
HisClα1	Fruit fly	F	W	I	K	P	E	A	A	P	A	R	V	T	L
GlyRβ	Mouse	F	W	I	N	P	D	A	S	A	A	R	V	P	L
$GABA_A$Rβ1	Mouse	F	W	I	N	Y	D	A	S	A	A	R	V	A	L
GluClβ	*C. elegans*	F	W	I	D	L	H	S	T	A	G	R	V	A	L
AChR B	*L. stagnalis*	F	W	L	P	P	E	S	-	P	A	K	M	Q	L
Capca1_92948	*C. capitata*	F	W	L	P	P	E	S	-	P	A	K	M	M	L
Smp_176310	*S. mansoni*	F	W	L	P	P	E	T	-	P	A	K	M	V	L
AChR2β	*S. haematobium*	F	W	L	P	P	D	S	-	A	A	K	M	M	L
Helro1_192409	*H. robusta*	F	L	L	P	P	D	T	-	N	A	K	I	S	L

b

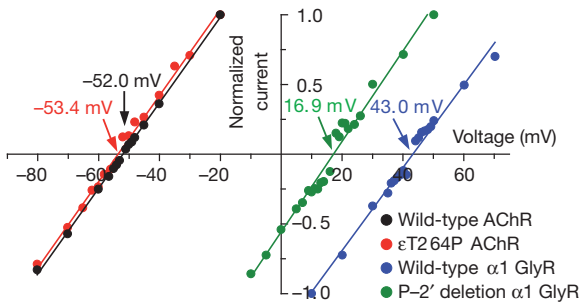


Figure 1 | The versatile charge selectivity of nicotinic-type receptors. **a**, Sequence alignment of residues in and flanking the M1–M2 loop. The broken horizontal line separates the sequences that are known or predicted (on the basis of their sequences) to form cation-selective channels (top) from those that are known or predicted to form anion-selective ones (bottom). Included in this alignment are subunits from receptors to acetylcholine (ACh), serotonin (5-HT), glycine (Gly), γ-aminobutyric acid (GABA), glutamate (Glu), histamine (His), and from receptors with as yet unidentified ligands. The invertebrate organisms in this list are: *Aplysia californica* (a mollusc), the fruit fly *Drosophila melanogaster* (an arthropod), *Caenorhabditis elegans* (a nematode), *Lymnaea stagnalis* (a mollusc), *Capitella capitata* (an annelid), *Schistosoma mansoni* and *S. haematobium* (two human parasitic platyhelminths) and *Helobdella robusta* (an annelid). **b**, Macroscopic current–voltage (*I*–*V*) relationships recorded under KCl-dilution conditions (solutions 8 and 9 in Supplementary Table 1; pH 7.4, both sides) in the outside-out configuration, as indicated in Supplementary Fig. 9 and in Methods. The equilibrium (Nernst) potentials at 22 °C, using ion concentration values, are −55.0 mV for K^+ and +50.6 mV for Cl^-. Reversal potentials are indicated. T264 denotes the threonine occupying position 12′, near the middle of M2, of the ε subunit of the AChR. P−2′ denotes the proline occupying position –2′ of the α1 subunit of the GlyR.

[1]Department of Molecular and Integrative Physiology, Center for Biophysics and Computational Biology, and Neuroscience Program, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.

level and frequent sojourns of brief duration in the zero-current ('shut') level (Fig. 2a; openings are downward deflections). This behaviour, which differs markedly from that of the wild-type channel under identical experimental conditions (Supplementary Fig. 2), suggested the occurrence of channel block by some of the components of the patch-bathing solutions. Indeed, omission of both Ca^{2+} and Mg^{2+} from the solution in the pipette eliminated this block and, at the same time, uncovered an unanticipated phenomenon involving two interconverting open-channel conductance levels (Fig. 2a). Unless otherwise indicated, all the results reported here correspond to recordings obtained in the absence of extracellular Ca^{2+} or Mg^{2+}.

As illustrated in Fig. 2b, c, the proline mutation in the four types of AChR subunit leads to the appearance of current fluctuations between two levels of open-channel current with the higher level (the 'main level') having roughly the same conductance as the single level observed in the wild-type channel, at least in the case of the δ- and ε-subunit mutants, where the fluctuations were most clearly resolved. Although the conductance of the lower level (the 'sublevel') and the occupancy probabilities of the two alternative open-channel states differ among mutants, the underlying phenomenon is undoubtedly the same, as expected from the nearly symmetrical arrangement of the five AChR M2 segments around the channel pore (Supplementary Fig. 3). Moreover, the kinetics of the fluctuations were found to depend on the pH of the external and internal solutions (Fig. 2d) in a manner that is fully consistent with these current oscillations reflecting the alternate protonation and deprotonation of an ionizable side chain. Thus, the effect of these proline mutations on ion conduction is highly reminiscent of the effect of lysine, arginine or histidine substitutions along the M1, M2 or M3 transmembrane segments of the muscle AChR[12,13]; the remarkable difference, however, is that the mutations reported here do not introduce any new protonatable group in the protein's amino-acid sequence.

To identify the residue(s) responsible for this phenomenon, we mutated each of the (native) ionizable amino acids in and flanking the M2 segment to non-ionizable residues while keeping the extra proline inserted between positions −2′ and −1′. Combining the results of mutations in the β1 and δ subunits (Fig. 3), we conclude that the observed main-level ⇌ sublevel transitions reflect the protonation and deprotonation of the 0′-lysine side chain of the subunit containing the proline mutation. Furthermore, as protonation reduces

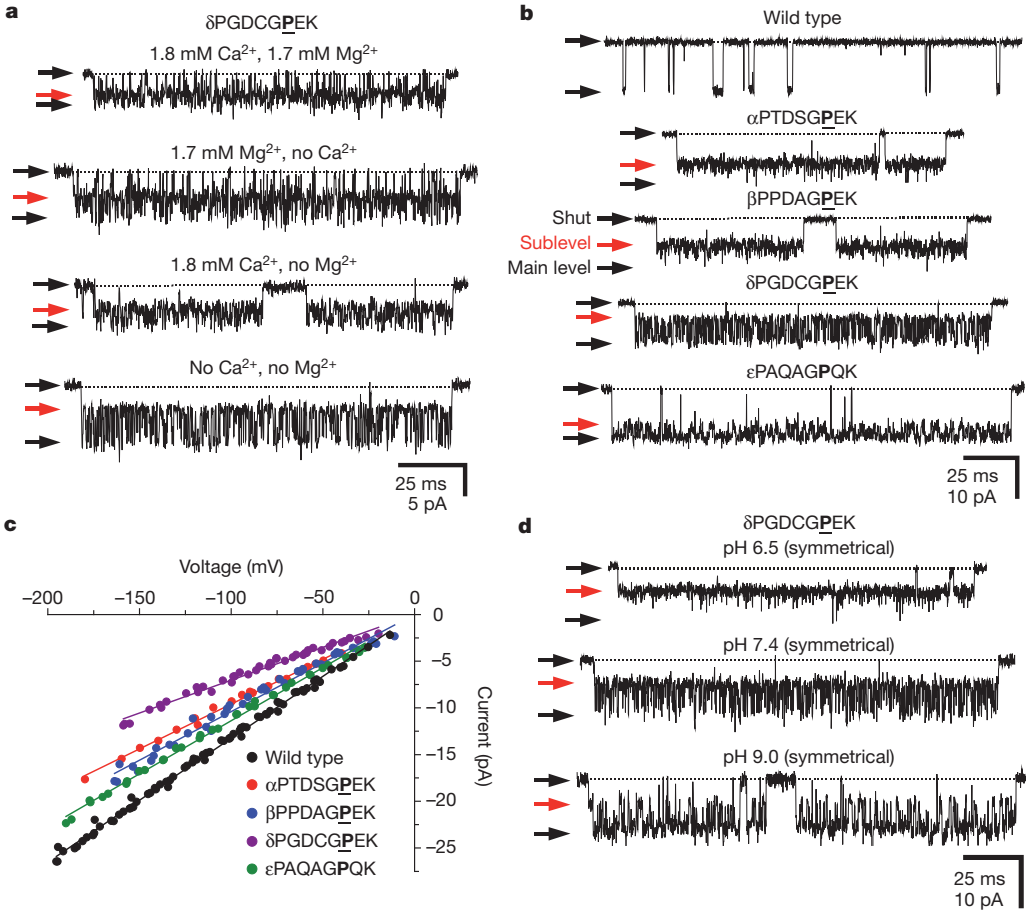


Figure 2 | A proline mutation unveils a proton-binding site. **a**, Single-channel inward currents (cell-attached configuration; approximately −100 mV; 1 μM ACh; pH of pipette solution ($pH_{pipette}$) 7.4) recorded from a mutant AChR having a proline inserted between positions −2′ and −1′ of the δ subunit. To increase the number of main-level ⇄ sublevel interconversions, a mutation that prolongs the mean duration of bursts of openings (εT264P) was also engineered. Solution compositions are indicated in Supplementary Table 1 (solutions 1–3). Mutations are indicated on the M1–M2 loop sequences; underlined bold symbols denote insertions whereas bold symbols (without the underline) denote substitutions. **b**, Inward currents (cell-attached configuration; approximately −100 mV; 1 μM ACh; $pH_{pipette}$ 7.4; solutions 2 and 3) recorded from the indicated AChR constructs. The burst-prolonging mutation was εT264P (in the case of AChRs with a proline inserted in the α1, β1 or δ subunit) or δS268Q (in the case of the glycine-to-proline substitution mutant at position −2′ of the ε subunit). In the case of the α1-subunit insertion, the trace shown corresponds to the construct having only one of the two α subunits mutated. **c**, Single-channel *I–V* relationships (cell-attached configuration; 1 μM ACh; $pH_{pipette}$ 7.4; solutions 2 and 3) recorded from the five constructs in **b**. For clarity, only the *I–V* curves corresponding to the sublevel are shown for the mutants. To facilitate the visual comparison of the slopes, each curve was displaced along the voltage axis so that it extrapolates exactly to the origin. **d**, pH dependence of the main-level ⇄ sublevel current fluctuations (outside-out configuration; −100 mV; 1 μM ACh; solutions 4 and 5).

the current amplitude (Fig. 2d), we also conclude that AChR mutants having a proline inserted (or substituted) into only one of the five subunits still conduct mostly cations.

A number of observations indicate that the lysines at position 0′ of the different AChR wild-type subunits reside on the stripe of M2 that faces away from the pore's lumen and that their εNH_2 groups are largely deprotonated, even at pH 6.0 (hence, $pK_a < 5.0$; Supplementary Text, Supplementary Figs 4, 5 and Supplementary Table 2). On introducing a proline, however, the affinity of these lysines for protons increases, probably as a result of a rearrangement of this portion of M2. This rearrangement does not seem to be a major one, however, because the extent to which the single-channel conductance is attenuated upon protonation of the 0′ lysines (a measure of the distance between the εNH_3^+ group and the long axis of the pore[12,13]; Supplementary Table 3) does not differ much from that caused by lysines engineered on the back or the sides of M2 (Supplementary Fig. 6) or the front of M1 or M3 (ref. 13). Moreover, although higher than in the wild-type AChR, the pK_a values of the εNH_3^+ group in the δ- and ε-subunit mutants (~7.58 and ~7.15, respectively; Supplementary Table 3) are still lower than the value expected for this group when fully exposed to bulk water (~10.4) by ~3 units (1 pK_a unit ≅ 1.36 kcal mol^{-1}). This further confirms the notion that these side chains do not face the aqueous lumen of the pore directly. For comparison, the pK_as of lysines engineered on the back of M2 (ref. 12) or on the front of M1 or M3 (ref. 13) are also, at least, ~3 units lower than the bulk-water value of ~10.4 (in Supplementary Table 4, we show that the pK_a values of substituted lysines are rather insensitive to the presence or absence of millimolar concentrations of external Ca^{2+} or Mg^{2+}). In addition, we conclude that the effect of these mutations is not highly position-specific because proline insertions at the five other possible positions along the M1–M2 loop of the δ subunit give rise, essentially, to the same pH-dependent phenotype (Supplementary Fig. 7).

The exact nature of the reorganization of the M1–M2 loop upon mutation, and how this change lowers the hydrophobicity of the microenvironment around the 0′ basic side chain, remains unknown. However, an increased exposure to water (through an increase in solvent penetration and/or a slight repositioning of the side chain) is expected to be an important factor in the stabilization of a positive charge buried in a region of the protein that lacks properly oriented acidic side chains. Indeed, note that the side chains of the nearby −5′ aspartate or the −1′ glutamate do not contribute to the observed main-level ⇄ sublevel current fluctuations (Fig. 3a), either by providing the proton-binding site itself or by electrostatically stabilizing the 0′ εNH_3^+ group. The idea of a structural rearrangement around position 0′ receives further support from the finding of a complex interaction between the proline mutants and extracellular Ca^{2+} and Mg^{2+} (compare Fig. 2a with Supplementary Fig. 2). Certainly, as elaborated in the Supplementary Text, it seems reasonable to ascribe the anomalous nature of this interaction to the probable concomitant rearrangement of the ring of glutamates at the neighbouring position −1′.

However, not all of the anion-selective members of the superfamily contain a full ring of 'inserted' prolines at position −2′ of the M1–M2 loop. Instead, some β-subunit homomers (such as those formed by the β subunits of $GABA_A$ receptors[14] or of invertebrate GluCl receptors[15]; Fig. 1a) present a full ring of alanines at this position without sacrificing high selectivity for anions. And, even more divergently, some highly anion-selective AChRs from invertebrates do not contain any extra residues in the M1–M2 loop, but rather replace the −2′ glycine of the cation-selective counterparts with a proline[16] (Fig. 1a). Remarkably, we found that mutating the muscle AChR to mimic the 'atypical' features of these M1–M2 loops also gives rise to current fluctuations that closely resemble those caused by the—more common—proline insertions characterized above. In fact, we found that the insertion or substitution of a number of amino acids (not only proline or alanine) in and around position −2′ have largely the same effect (Fig. 4 and Supplementary Text).

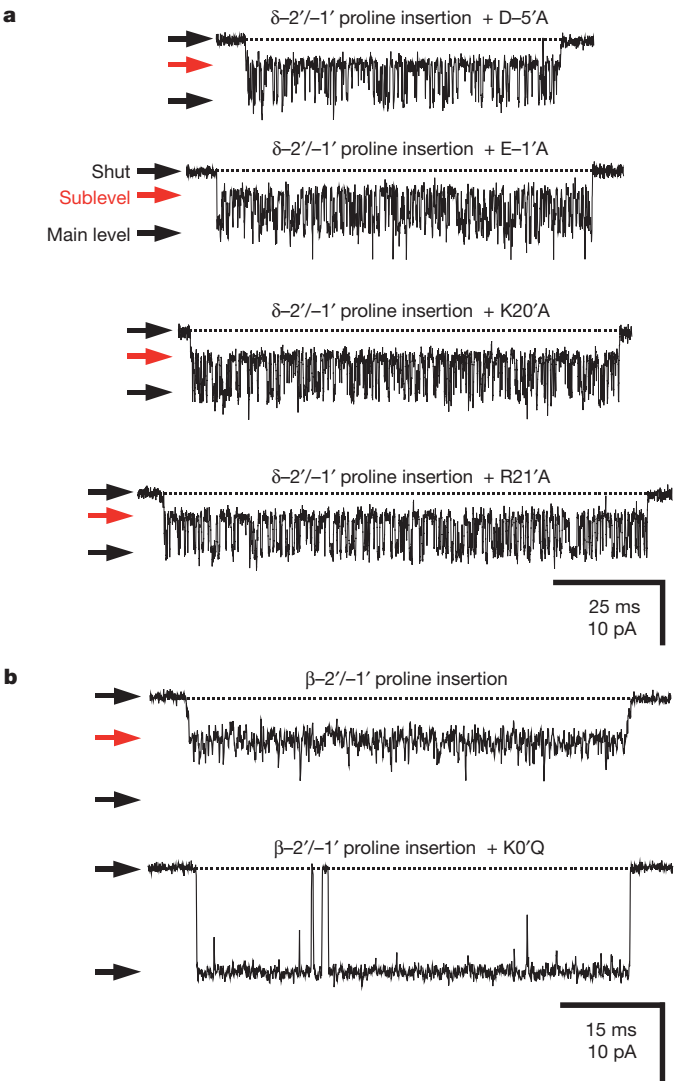


Figure 3 | The side chain of the 0′ basic residue is the proton-binding site. **a**, Single-channel inward currents (cell-attached configuration; 1 μM ACh; $pH_{pipette}$ 7.4; solutions 2 and 3) recorded from AChRs with a proline inserted between positions −2′ and −1′ of the δ subunit and having four of the five native ionizable residues that flank δM2 mutated to alanine, one at a time. The burst-prolonging mutation was εT264P. Mutation of the fifth residue (the 0′ lysine) to alanine, glutamine or valine (in the presence of the inserted proline) abolishes receptor expression in the plasma membrane, as revealed by the lack of specific α-bungarotoxin binding. The applied potential was approximately −100 mV for all constructs, with the exception of the receptor containing the glutamate-to-alanine mutation at position −1′, in which case the potential was approximately −150 mV (to compensate for its lower single-channel conductance). **b**, Inward currents recorded from a mutant AChR having a proline inserted between positions −2′ and −1′ of the β1 subunit and from the mutant having, in addition, a lysine-to-glutamine mutation at position 0′ of the same subunit. The applied potential was approximately −100 mV. All other experimental conditions were as in **a**.

Evidently, charge-selective permeation through members of the nicotinic-receptor superfamily has arisen during evolution as a result of several different amino-acid changes in the M1–M2 loop; yet, all these changes seem to act, at least in part, by tuning the proton affinity of the same basic side chain. Consistent with the functional relevance of the different protonation states of the ionizable group at 0′ (neutral in cation-selective members and, at least partly, positively charged in anion-selective members) we notice that some recently identified cation-selective members of the superfamily from nematodes[17] and bacteria[18] replace this lysine or arginine with non-ionizable residues. However, all known native anion-selective nicotinic-type receptors

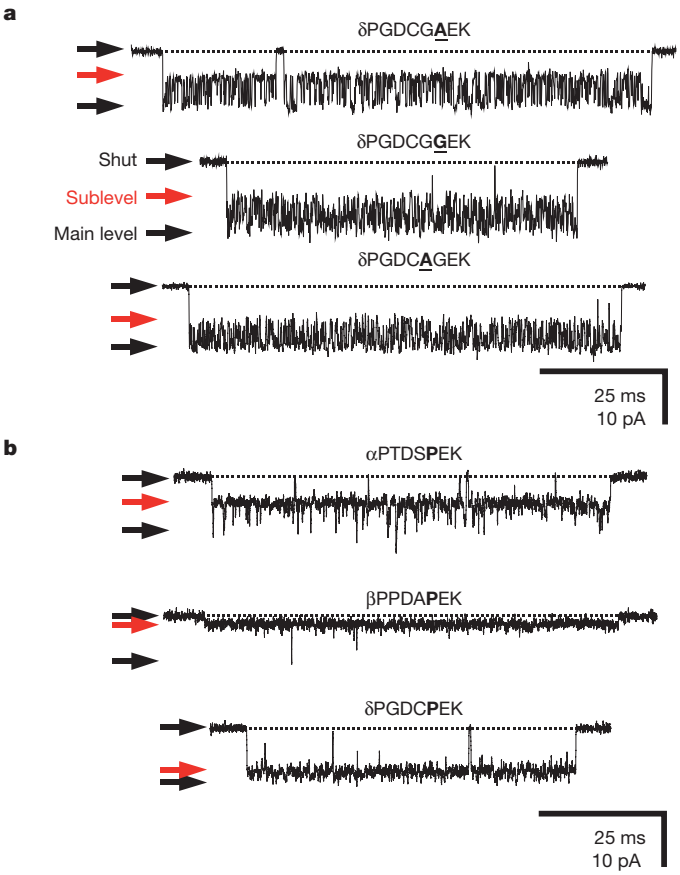


Figure 4 | Not only prolines, not only insertions. **a**, Single-channel inward currents (cell-attached configuration; approximately −100 mV; 1 μM ACh; $pH_{pipette}$ 7.4; solutions 2 and 3) recorded from the indicated AChR insertion mutants. The burst-prolonging mutation was εT264P. Threonine insertions have a similar effect. **b**, Inward currents recorded from the indicated AChR substitution mutants under the same experimental conditions as in **a**. Note that the insertion of a residue is not required to reveal a proton-binding site. Instead, replacing the conserved glycine at position −2′ with a variety of other residues (see Supplementary Text; only proline is shown here) also unveils a protonation site in the four types of subunit. In the case of the α1-subunit mutant, the trace shown corresponds to the construct having only one of the two α subunits mutated. The trace illustrating the effect of a glycine-to-proline mutation at this position of the ε subunit is shown in Fig. 2b.

present a basic residue at position 0′. Whether these additional positive charges are directly responsible for the anion selectivity or, rather, they act to increase the single-channel current amplitude of a channel that is highly selective for anions irrespective of the protonation state of the 0′ basic side chains (as a result, perhaps, of concomitant changes in pore size[6,11,19] or in the orientation of backbone groups[2,10]) remains unclear. What is clear, however, is that both high charge selectivity and high single-channel current amplitude are essential for proper electrical signalling at fast chemical synapses. What is also clear is that the differential tuning of side-chain pK_a values described here represents a novel mechanism for turning protein charges on or off without the need of replacing ionizable amino acids with non-ionizable ones (or vice versa).

It is worth noting that the finding of different protonation states for the 0′ basic side chain in cation-selective-type versus anion-selective-type charge-selectivity filters would have gone unnoticed by even such powerful approaches as X-ray or electron crystallography. Certainly, these methods do not typically reach the resolution of 1.0–1.2 Å (especially when applied to membrane proteins) that is needed to detect the presence of hydrogen atoms. Also, although a variety of structure-based computational algorithms for the prediction of protein side-chain pK_a values have been developed and could in principle be applied to structural models of members of the nicotinic-receptor superfamily, their accuracy in the case of large deviations from values in bulk water is still very limited[20–24].

Overall, our results provide a compelling example of the marked sensitivity of side-chain pK_a values to the details of the microenvironment, of the profound impact that differentially tuned proton affinities can have on protein function, and of the advantage evolution has taken of this physicochemical phenomenon. Lastly, our data also remind us that assuming default protonation states for the ionizable side chains in a protein may be highly misleading, and that the structural determinants of ion-conduction properties through ion channels need not face the lumen of the pore directly.

METHODS SUMMARY

Currents were recorded from HEK-293 cells transiently transfected with wild-type or mutant complementary DNAs (cDNAs) encoding the adult muscle-type AChR (mouse α1, β1, δ and ε subunits) or the α1 GlyR (human or rat isoform b). Single-channel currents were recorded at 22 °C from cell-attached patches with the exception of recordings that required access to both sides of the membrane in which case the outside-out configuration with a constant application of ligand was used. Ensemble ('macroscopic') currents were recorded at 22 °C from outside-out patches exposed to step changes in the concentration of ligand. The composition of all solutions used for electrophysiological recordings is given in Supplementary Table 1. Extent-of-channel-block and pK_a values were estimated from cell-attached, single-channel recordings as detailed in our previous work[12,13] and in Supplementary Fig. 8. All single-channel current traces are displayed at $f_c \approx 6$ kHz. Reversal potentials were estimated from macroscopic-current recordings elicited by 1- or 10-ms pulses of ligand applied to outside-out patches (Supplementary Fig. 9). The expression of mutant AChRs in the plasma membrane of transfected cells was estimated using an equilibrium [^{125}I]-α-bungarotoxin binding assay.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 5 April 2010; accepted 21 March 2011.
Published online 22 May 2011.

1. Galzi, J. L. *et al.* Mutations in the ion channel domain of a neuronal nicotinic receptor convert ion selectivity from cationic to anionic. *Nature* **359,** 500–505 (1992).
2. Corringer, P.-J. *et al.* Mutational analysis of the charge selectivity filter of the α7 nicotinic acetylcholine receptor. *Neuron* **22,** 831–843 (1999).
3. Keramidas, A., Moorhouse, A. J., French, C. R., Schofield, P. R. & Barry, P. H. M2 pore mutations convert the glycine receptor channel from being anion- to cation-selective. *Biophys. J.* **79,** 247–259 (2000).
4. Gunthorpe, M. J. & Lummis, S. C. R. Conversion of the ion selectivity of the 5-HT_{3A} receptor from cationic to anionic reveals a conserved feature of the ligand-gated ion channel superfamily. *J. Biol. Chem.* **276,** 10977–10983 (2001).
5. Jensen, M. L. *et al.* The β subunit determines the ion selectivity of the $GABA_A$ receptor. *J. Biol. Chem.* **277,** 41438–41447 (2002).
6. Keramidas, A., Moorhouse, A. J., Pierce, K. D., Schofield, P. R. & Barry, P. H. Cation-selective mutations in the M2 domain of the inhibitory glycine receptor channel reveal determinants of ion-charge selectivity. *J. Gen. Physiol.* **119,** 393–410 (2002).
7. Thompson, A. J. & Lummis, S. C. R. A single ring of charged amino acids at one end of the pore can control ion selectivity in the 5-HT_3 receptor. *Br. J. Pharmacol.* **140,** 359–365 (2003).
8. Wotring, V. E., Miller, T. S. & Weiss, D. S. Mutations at the GABA receptor selectivity filter: a possible role for effective charges. *J. Physiol. (Lond.)* **548,** 527–540 (2003).
9. Menard, C., Horvitz, H. R. & Cannon, S. Chimeric mutations in the M2 segment of the 5-hydroxytryptamine-gated chloride channel MOD-1 define a minimal determinant of anion/cation permeability. *J. Biol. Chem.* **280,** 27502–27507 (2005).
10. Sunesen, M. *et al.* Mechanism of Cl^- selection by a glutamate-gated chloride (GluCl) receptor revealed through mutations in the selectivity filter. *J. Biol. Chem.* **281,** 14875–14881 (2006).
11. Keramidas, A., Moorhouse, A. J., Schofield, P. R. & Barry, P. H. Ligand-gated ion channels: mechanisms underlying ion selectivity. *Prog. Biophys. Mol. Biol.* **86,** 161–204 (2004).
12. Cymes, G. D., Ni, Y. & Grosman, C. Probing ion-channel pores one proton at a time. *Nature* **438,** 975–980 (2005).
13. Cymes, G. D. & Grosman, C. Pore-opening mechanism of the nicotinic acetylcholine receptor evinced by proton transfer. *Nature Struct. Mol. Biol.* **15,** 389–396 (2008).
14. Krishek, B. J., Moss, S. J. & Smart, T. G. Homomeric β1 γ-aminobutyric $acid_A$ receptor-ion channels: evaluation of pharmacological and physiological properties. *Mol. Pharmacol.* **49,** 494–504 (1996).
15. Cully, D. F. *et al.* Cloning of an avermectin-sensitive glutamate-gated chloride channel from *Caenorhabditis elegans*. *Nature* **371,** 707–711 (1994).

16. van Nierop, P. *et al.* Identification of molluscan nicotinic acetylcholine receptor (nAChR) subunits involved in formation of cation- and anion-selective nAChRs. *J. Neurosci.* **25,** 10617–10626 (2005).
17. Beg, A. A. & Jorgensen, E. M. EXP-1 is an excitatory GABA-gated cation channel. *Nature Neurosci.* **6,** 1145–1152 (2003).
18. Bocquet, N. *et al.* A prokaryotic proton-gated ion channel from the nicotinic acetylcholine receptor family. *Nature* **445,** 116–119 (2007).
19. Lee, D. J.-S., Keramidas, A., Moorhouse, A. J., Schofield, P. R. & Barry, P. H. The contribution of proline 250 (P-2′) to pore diameter and ion selectivity in the human glycine receptor channel. *Neurosci. Lett.* **351,** 196–200 (2003).
20. Schutz, C. N. & Warshel, A. What are the dielectric 'constants' of proteins and how to validate electrostatic models? *Proteins* **44,** 400–417 (2001).
21. Harms, M. J. *et al.* The pK_a values of acidic and basic residues buried at the same internal location in a protein are governed by different factors. *J. Mol. Biol.* **389,** 34–47 (2009).
22. Kamerlin, S. C. L., Haranczyk, M. & Warshel, A. Progress in *ab initio* QM/MM free-energy simulations of electrostatic energies in proteins: accelerated QM/MM studies of p*K*, redox reactions and solvation free energies. *J. Phys. Chem. B* **113,** 1253–1272 (2009).
23. Karp, D. A., Stahley, M. R. & García-Moreno, E. B. Conformational consequences of ionization of Lys, Asp, and Glu buried at position 66 in staphylococcal nuclease. *Biochemistry* **49,** 4138–4146 (2010).
24. Chimenti, M. S., Castañeda, C. A., Majumdar, A. & García-Moreno, E. B. Structural origins of high apparent dielectric constants experienced by ionizable groups in the hydrophobic core of a protein. *J. Mol. Biol.* **405,** 361–377 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Sine for wild-type muscle AChR cDNA; M. Slaughter and D. Papke for wild-type α1 GlyR cDNA; S. Elenes for critical advice on fast-perfusion experiments; E. Jakobsson and H. Robertson for discussions; and G. Papke, M. Maybaum, J. Pizarek and C. Staehlin for technical assistance. This work was supported by a grant from the US National Institutes of Health (R01-NS042169 to C.G.).

Author Contributions G.D.C. and C.G. designed experiments, analysed data and wrote the manuscript; G.D.C performed experiments.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to C.G. (grosman@illinois.edu).

METHODS

DNA clones, mutagenesis and transfection. HEK-293 cells were transiently transfected with cDNAs encoding the adult muscle-type AChR (mouse α1, β1, δ and ε subunits) or the α1 GlyR (human or rat isoform b prepared as indicated in ref. 25; no differences were found between the charge selectivities of these two orthologues) using a calcium-phosphate precipitation method. Mutations were engineered using the QuikChange site-directed mutagenesis kit (Stratagene) and were confirmed by dideoxy sequencing. When deemed necessary, mutations that prolong individual activations of the channel ('bursts of openings') were also introduced in the mutant AChR constructs to increase the number of proton-transfer events recorded. These mutations were βV266M (M2 position 13′; ref. 26), δS268Q (M2 12′; ref. 27) or εT264P (M2 12′; ref. 28), and their lack of appreciable effect on charge selectivity and single-channel conductance is shown in Fig. 1b and Supplementary Fig. 10, respectively.

Extent of channel block. Single-channel currents were digitized (at 100 kHz), filtered (cascaded $f_c \approx 30$ kHz) and idealized (using the SKM algorithm in QuB software[29]) to obtain the mean amplitudes of the different current levels and the sequences of dwell times. *I*–*V* curves were generated from patch-clamp recordings obtained in the cell-attached configuration, and these only include data on inward currents (for example, Fig. 2c). For most mutants studied here, the rectilinear portion of the sublevel's *I*–*V* curve extrapolates onto the voltage axis at a negative potential (whereas the main-level's curve extrapolates near zero), a probable result of the more pronounced inward rectification of the current sublevel and, for some mutants at least, the result perhaps of the diminished selectivity of the sublevel for cations (note that, under the conditions of our cell-attached experiments, a decrease in cation selectivity would shift the reversal potential to negative values). Because of these different intercepts, the ratio between the sublevel and the main-level single-channel current amplitudes becomes a function of the transmembrane potential, and the choice of any particular voltage value to calculate the extent of channel block from current amplitudes would be arbitrary. Hence, here (as in our previous work[12,13]) we chose to calculate the extent of block using single-channel conductances, instead. As a result, our extent-of-block values differ from those that could be inferred from a mere inspection of the single-channel traces at a single potential (for example, −100 mV, in the case of our figures). Different intercepts for the rectilinear portions of the main-level and sublevel *I*–*V* curves are not unique to the mutants studied here; rather, these differences were also observed for AChR mutants bearing engineered basic residues along M2 (ref. 12). The extent of channel block for each construct was calculated as the difference between the conductance values of the main level and the sublevel normalized by the conductance of the main level. In some cases, the conductance of the main level could not be estimated with confidence (for example, because the open-channel signal dwelled only briefly and infrequently in the main level). In these cases, the normalization was done relative to the conductance of the corresponding background construct (that is, the wild-type AChR with or without one of the burst-prolonging mutations).

pK_a values. Protonation and deprotonation rates (as well as all other transition rates) were estimated from maximum-likelihood fits of single-channel dwell-time sequences with kinetic models (Supplementary Fig. 8) as described in our previous work[12,13]. To this end, we used the MIL algorithm in QuB software[30] with a retrospectively imposed time resolution of 25 μs. The ratio between the proton-dissociation and proton-association rates thus estimated gives the ratio of the probabilities of the engineered ionizable side chain being deprotonated versus protonated while the channel is open. The reported pK_a values (Supplementary Tables 3 and 4) were calculated from the product of these ratios and the concentration of protons in the channel-bathing solution (Supplementary Fig. 8). For the calculation of the pK_as of lysine side chains engineered in M1 or M2 (Supplementary Table 4), we used the concentration of protons in the pipette solution of cell-attached patches (that is, pH 6.0 in the case of the mutant at position 11′ and 7.4 in all other cases). For the calculation of the pK_as of the 0′ side chain in the various M1–M2 loop mutants studied here (Supplementary Table 3), however, the choice of a pH value is not obvious because we found that the kinetics of protonation and deprotonation in this region of the channel are sensitive to the pHs of the two solutions bathing the membrane, behaving as if the protonatable group were exposed to a solution of intermediate pH. Hence, although probably not strictly correct, we decided to use a pH of 7.3, a value halfway between the pH of the pipette solution (~7.4) and that of the cytosol (~7.2). This uncertainty leads to a maximum systematic error of ± 0.1 units in the pK_a estimates shown in Supplementary Table 3.

Concentration jumps, reversal potentials and kinetics. Step changes in the concentration of ligand bathing the external aspect of outside-out patches were achieved by the rapid switching of two solutions (differing only in the presence or absence of ligand) flowing from either barrel of a piece of theta-type capillary glass mounted on a piezo-electric device (Burleigh-LSS-3100; Lumen Dynamics) as described previously[31] (solution-exchange time$_{10–90\%}$ < 150 μs). Reversal potentials were estimated from *I*–*V* relationships generated by plotting the peak-current responses to brief (1- or 10-ms) pulses of ligand applied to outside-out patches at concentrations that evoke nearly maximal responses (100 μM ACh for the AChR; 10 mM Gly for the α1 GlyR). Consecutive pulses were separated by 8-s intervals during which the patches were exposed to ligand-free solution and the applied voltage was changed. In these particular experiments, the reference Ag/AgCl wire was connected to the bath solution (the composition of which was the same as that of the solution flowing through the theta-type glass tubing; solution 9 in Supplementary Table 1) through an agar bridge containing 200 mM KCl, to minimize the liquid-junction potential. A new, fresh agar bridge was connected every <2 h. Liquid-junction potentials were calculated using the JPCalc module in pClamp 9.0 (ref. 32). To characterize the kinetics of AChR deactivation, entry into desensitization and recovery from desensitization, and the response to the repetitive application (25 Hz) of nearly saturating ACh, macroscopic currents were recorded from outside-out patches (at –80 mV) using various 100-μM ACh pulse protocols, as indicated in Supplementary Fig. 5. All macroscopic currents were analysed using a combination of pClamp 9.0 (Molecular Devices) and SigmaPlot 7.101 (Systat Software) software.

Plasma-membrane AChR expression. To estimate the number of wild-type or mutant AChRs in the plasma membrane, transfected HEK-293 cells were incubated with 20-nM [^{125}I]-α-bungarotoxin (PerkinElmer) in fresh DMEM culture medium at 4–5 °C for 2–3 h so as to saturate all toxin-binding sites. The associated radioactivity was measured in a γ-counter and was normalized to the corresponding mass of total protein, which was quantified using the bicinchoninic-acid method (Thermo Scientific) after solubilizing the cells with 0.1 N NaOH. The non-specific binding of radiolabelled toxin was estimated on cells transfected with cDNA encoding the β1, δ and ε subunits of the mouse-muscle AChR (but not the α1 subunit). The amount of [^{125}I]-α-bungarotoxin bound to these mock-transfected cells (normalized to total protein content) was never higher than 6% of that associated with the expression of the wild-type AChR.

25. Papke, D., Gonzalez-Gutierrez, G. & Grosman, C. Desensitization of neurotransmitter-gated ion channels during high-frequency stimulation: a comparative study of Cys-loop, AMPA and purinergic receptors. *J. Physiol. (Lond.)* **589,** 1571–1585 (2011).
26. Engel, A. G. *et al.* New mutations in acetylcholine receptor subunit genes reveal heterogeneity in the slow-channel congenital myasthenic syndrome. *Hum. Mol. Genet.* **5,** 1217–1227 (1996).
27. Grosman, C. & Auerbach, A. Asymmetric and independent contribution of the second transmembrane segment 12′ residues to diliganded gating of acetylcholine receptor channels. A single-channel study with choline as the agonist. *J. Gen. Physiol.* **115,** 637–651 (2000).
28. Ohno, K. *et al.* Congenital myasthenic syndrome caused by prolonged acetylcholine receptor channel openings due to a mutation in the M2 domain of the epsilon subunit. *Proc. Natl Acad. Sci. USA* **92,** 758–762 (1995).
29. Qin, F. Restoration of single-channel currents using the segmental k-means method based on hidden Markov modeling. *Biophys. J.* **86,** 1488–1501 (2004).
30. Qin, F., Auerbach, A. & Sachs, F. Estimating single-channel kinetic parameters from idealized patch-clamp data containing missed events. *Biophys. J.* **70,** 264–280 (1996).
31. Elenes, S., Ni, Y., Cymes, G. D. & Grosman, C. Desensitization contributes to the synaptic response of gain-of-function mutants of the muscle nicotinic receptor. *J. Gen. Physiol.* **128,** 615–627 (2006).
32. Barry, P. H. & Lynch, J. W. Liquid junction potentials and small cell effects in patch-clamp analysis. *J. Membr. Biol.* **121,** 101–117 (1991).

LETTER

doi:10.1038/nature10072

Detection of prokaryotic mRNA signifies microbial viability and promotes immunity

Leif E. Sander[1], Michael J. Davis[2]*, Mark V. Boekschoten[3]*, Derk Amsen[4], Christopher C. Dascher[1], Bernard Ryffel[5], Joel A. Swanson[2], Michael Müller[3] & J. Magarian Blander[1]

Live vaccines have long been known to trigger far more vigorous immune responses than their killed counterparts[1–6]. This has been attributed to the ability of live microorganisms to replicate and express specialized virulence factors that facilitate invasion and infection of their hosts[7]. However, protective immunization can often be achieved with a single injection of live, but not dead, attenuated microorganisms stripped of their virulence factors. Pathogen-associated molecular patterns (PAMPs), which are detected by the immune system[8,9], are present in both live and killed vaccines, indicating that certain poorly characterized aspects of live microorganisms, not incorporated in dead vaccines, are particularly effective at inducing protective immunity. Here we show that the mammalian innate immune system can directly sense microbial viability through detection of a special class of viability-associated PAMPs (vita-PAMPs). We identify prokaryotic messenger RNA as a vita-PAMP present only in viable bacteria, the recognition of which elicits a unique innate response and a robust adaptive antibody response. Notably, the innate response evoked by viability and prokaryotic mRNA was thus far considered to be reserved for pathogenic bacteria, but we show that even non-pathogenic bacteria in sterile tissues can trigger similar responses, provided that they are alive. Thus, the immune system actively gauges the infectious risk by searching PAMPs for signatures of microbial life and thus infectivity. Detection of vita-PAMPs triggers a state of alert not warranted for dead bacteria. Vaccine formulations that incorporate vita-PAMPs could thus combine the superior protection of live vaccines with the safety of dead vaccines.

We hypothesized that the innate immune system might sense the most fundamental characteristic of microbial infectivity, microbial viability itself, and activate a robust immune response regardless of the presence of more specialized factors that regulate microbial virulence[7]. To study the sensing of bacterial viability without the compounding effects of replication or virulence factors, we used thymidine auxotrophs of non-pathogenic *Escherichia coli* K12, strain DH5α (hereafter called *thyA*$^-$ *E. coli*). Viable and heat-killed *thyA*$^-$ *E. coli* similarly activated nuclear factor-κB (NF-κB) and mitogen-activated protein kinase p38 (Supplementary Fig. 1) in murine bone-marrow-derived macrophages and elicited production of similar amounts of interleukin-6 (IL-6) and tumour necrosis factor-α (TNF-α) (Fig. 1a). In contrast, viable *thyA*$^-$ *E. coli* induced higher levels of IFN-β than heat-killed *thyA*$^-$ *E. coli* or lipopolysaccharide (LPS) (Fig. 1b), and only viable *thyA*$^-$ *E. coli* induced IL-1β secretion (Fig. 1c and Supplementary Fig. 2). Pro-IL-1β transcription was equally induced by both viable and heat-killed *thyA*$^-$ *E. coli* (Fig. 1c), indicating that viable bacteria specifically elicit cleavage of pro-IL-1β. This process is catalysed by caspase-1 in Nod-like receptor (NLR)-containing inflammasome complexes, the assembly of which can be triggered by the activity of bacterial virulence factors[10,11]. Notably, avirulent viable but not heat-killed *thyA*$^-$ *E. coli* induced inflammasome activation and pro-caspase-1 cleavage (Fig. 1d). Finally, viable but not heat-killed *thyA*$^-$ *E. coli* induced caspase-1-dependent inflammatory cell death, termed pyroptosis[10,11], resulting in the release of lactate dehydrogenase (LDH) (Fig. 1e) and the appearance of 7-amino-actinomycin D (7AAD)$^+$annexin-V$^{-/low}$ cells (Fig. 1f). Similar responses were observed in peritoneal macrophages and both splenic and bone-marrow-derived dendritic cells (Supplementary Fig. 2b). Killing *thyA*$^-$ *E. coli* by ultraviolet irradiation, antibiotics, or ethanol also selectively abrogated IL-1β secretion and pyroptosis without affecting IL-6 production (Fig. 1g and Supplementary Fig. 3), indicating that a general determinant associated with bacterial viability is detected.

To determine whether pathogenic bacteria can also activate the inflammasome in the absence of virulence factors, we studied attenuated strains of selected pathogens: *Shigella flexneri* virulence plasmid-cured strain BS103[12], *Salmonella enterica* serovar Typhimurium SL1344 Δ*Spi1*Δ*Spi2*, lacking the *Salmonella* pathogenicity islands SPI-1 and SPI-2 (ref. 10), and *Listeria monocytogenes* Δ*Hly*Δ*fliC*, lacking listeriolysin O and flagellin[10]. These mutants induced IL-1β production at levels comparable to those induced by *thyA*$^-$ *E. coli* (Fig. 1h), but lower and with slower kinetics than their pathogenic counterparts (Supplementary Fig. 4). IL-1β production was abolished when these bacteria were killed, whereas IL-6 production was similar (Fig. 1h). Thus, immune cells detect universal characteristics of viability different from virulence factors.

Caspase-1 activation, pyroptosis and IL-1β production in response to *thyA*$^-$ *E. coli* were abrogated in macrophages deficient for NLRP3 or for the inflammasome adaptor apoptosis speck protein with caspase recruitment (ASC or PYCARD)[11] (Fig. 1i, j), whereas NLRC4 was dispensable (Supplementary Fig. 5). Pyroptosis and IL-1β production induced by viable *thyA*$^-$ *E. coli* were abrogated in *Casp1*$^{-/-}$ macrophages (Fig. 1j) and suppressed by inhibitors for caspase-1, but not caspase-8 (Supplementary Fig. 6).

Induction of IFN-β mRNA and protein by viable *thyA*$^-$ *E. coli* required the Toll-like receptor (TLR) adaptor TRIF[9] (Fig. 2a, b) and downstream interferon regulatory factor-3 (IRF3)[9] (Supplementary Fig. 7), but not MyD88, the main TLR adaptor[9] (Fig. 2a, b). In contrast, transcription of pro-IL-1β was largely dependent on MyD88. Consequently, *Myd88*$^{-/-}$ cells secreted no IL-1β (Fig. 2c, d), whereas pyroptosis and caspase-1 cleavage were intact (Fig. 2e, f). Notably, although TRIF was dispensable for pro-IL-1β transcription (Fig. 2c), *Trif*$^{-/-}$ cells failed to secrete IL-1β (Fig. 2d), were protected from pyroptosis (Fig. 2e) and did not activate caspase-1 (Fig. 2f). These findings revealed an unexpected role for TRIF in NLRP3 inflammasome activation in response to viable *thyA*$^-$ *E. coli*. In contrast, pyroptosis induced by pathogenic *S. enterica* Typhimurium[10] proceeded independently of TRIF (Supplementary Fig. 8). Differential involvement of TRIF, together with differences in magnitude and kinetics of the response (Fig. 1h and Supplementary Fig. 4), indicated that inflammasome activation in

[1]Immunology Institute, Department of Medicine, Mount Sinai School of Medicine, 1425 Madison Avenue, New York, New York 10029, USA. [2]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan 48109-0620, USA. [3]Nutrition, Metabolism and Genomics Group, Division of Human Nutrition, Wageningen University, 6703 HD Wageningen, The Netherlands. [4]Department of Cell Biology and Histology, Academic Medical Center, University of Amsterdam, 1105 AZ Amsterdam, The Netherlands. [5]Laboratory of Molecular Immunology and Embryology, University of Orleans and Centre National de la Recherche Scientifique, 45071 Orleans, France.

*These authors contributed equally to this work.

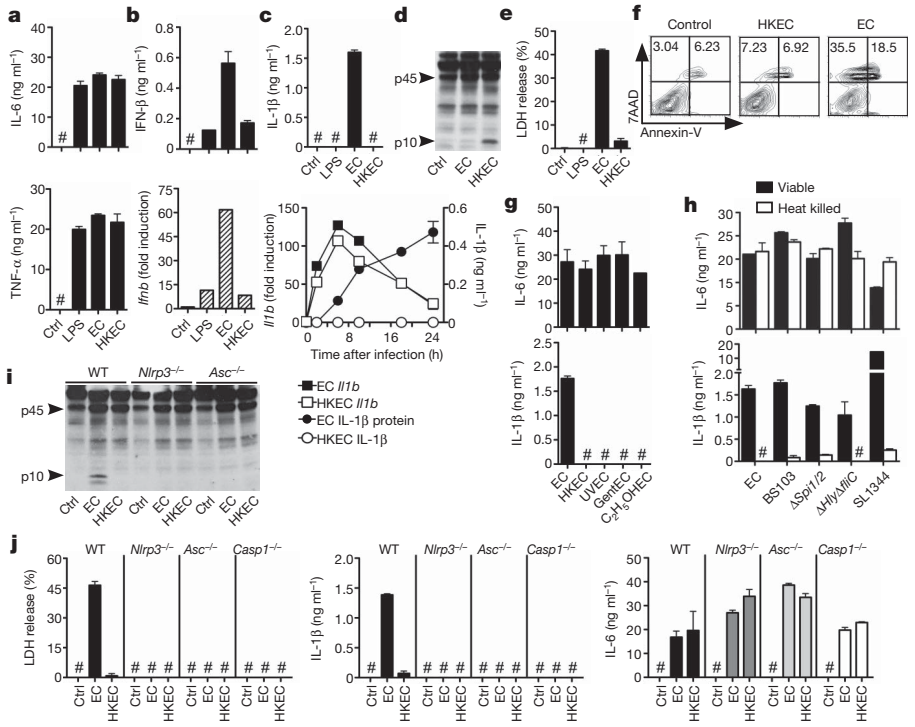


Figure 1 | Sensing bacterial viability induces IFN-β and activates the NLRP3 inflammasome in the absence of virulence factors. **a**, **b**, IL-6 and TNF-α (**a**) and IFN-β protein and mRNA (at 2 h) (**b**) levels in murine BMMs stimulated with medium (ctrl), lipopolysaccharide (LPS), *thyA*$^-$ *E. coli* (EC) and heat-killed *thyA*$^-$ *E. coli* (HKEC). Multiplicity of infection = 20. **c**, IL-1β (top), *Il1b* mRNA (bottom, left *y* axis) and secreted IL-1β (bottom, right *y* axis) at indicated times is shown. **d**, **i**, Caspase-1 immunoblots at 18 h in wild-type (**d**) or wild-type (WT), *Nlrp3*$^{-/-}$ and *Asc*$^{-/-}$ BMMs (**i**). **e**, **f**, Pyroptosis by LDH release (**e**) and FACS (**f**) at 18 h is shown. **g**, **h**, IL-6 and IL-1β in response to *thyA*$^-$ *E. coli*, viable or killed by different means (**g**, bone-marrow-derived dendritic cells (BMDCs)) or viable or heat-killed *thyA*$^-$ *E. coli*, attenuated *thyA*$^-$ *Shigella* (BS103), *Salmonella* (Δ*Spi1/2*) and *Listeria* (Δ*Hly*Δ*fliC*), or virulent *Salmonella* SL1344 (**h**). C_2H_5OHEC, ethanol-killed *E. coli*; GentEC, gentamicin-killed *E. coli*; UVEC, UV irradiated *E. coli*. **j**, LDH, IL-1β and IL-6 in BMMs of the indicated genotype in response to medium (ctrl), viable *thyA*$^-$ *E. coli* (EC) and heat-killed *thyA*$^-$ *E. coli* (HKEC). All responses are by murine BMMs and measured at 24 h unless indicated otherwise. Hash symbol indicates not detected. Data represent ≥5 experiments. All bars represent mean ± s.e.m.

response to virulence factors occurs in a manner distinct from that to viability.

Genome-wide transcriptional analysis of wild-type and *Trif*$^{-/-}$ macrophages before and after phagocytosis of viable *thyA*$^-$ *E. coli* showed differential regulation of several clusters of genes (Supplementary Fig. 9) including IFN-regulated genes, as expected[9] (Fig. 2a, b and Supplementary Fig. 10a), whereas most of the Rel/NF-κB target genes were comparable (Supplementary Fig. 10b). *Nlrp3* expression was induced independently of TRIF (Fig. 2g, h), and negative regulators of inflammasome activity, such as those encoded by Mediterranean fever (*Mefv*), *Nlrp10* and *Casp12* genes, were also unchanged or expressed at higher levels in wild-type macrophages (Fig. 2g), possibly due to negative feedback. Thus, the role of TRIF in inflammasome activation upon phagocytosis of viable *thyA*$^-$ *E. coli* is not explained by transcriptional control of inflammasome components (so called priming[11]). Furthermore, ATP and reactive oxygen species (ROS)[11,13], known activators of the NLRP3 inflammasome, were not involved, as deficiency for P_2X_7R, which is required for ATP-mediated NLRP3 activation, did not affect pyroptosis or IL-1β production (Supplementary Fig. 11a, b), and ROS accumulated equally in response to viable and heat-killed *thyA*$^-$ *E. coli* independently of TRIF (Supplementary Fig. 11c).

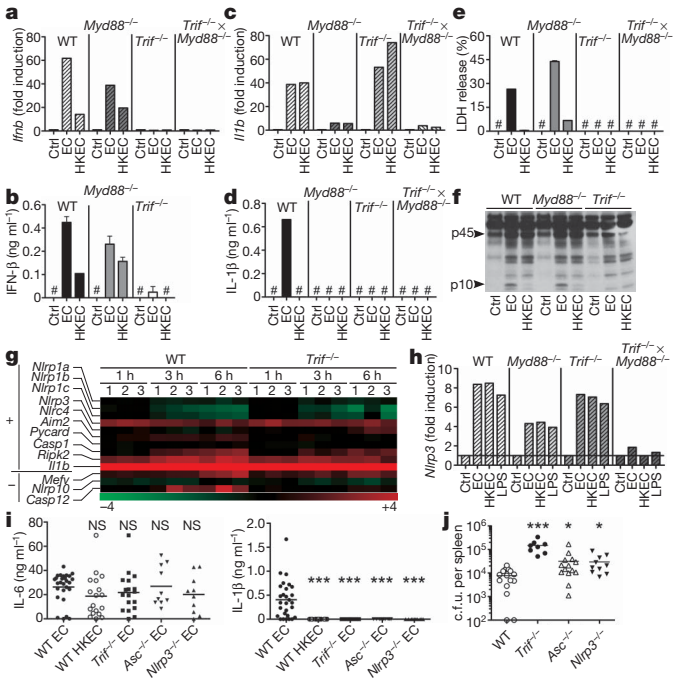


Figure 2 | The TLR signalling adaptor TRIF controls 'viability-induced' responses. **a**–**e**, *Ifnb* transcription at 2 h (**a**), IFN-β secretion at 24 h (**b**), *Il1b* transcription at 2 h (**c**), IL-1β secretion (**d**) and LDH release (**e**) at 24 h after phagocytosis of viable (EC) or heat-killed (HKEC) *thyA*$^-$ *E. coli*. **f**, Caspase-1 immunoblot at 18 h. Data in **a**–**f** are from murine BMMs and represent ≥5 experiments. **g**, Gene microarray analysis of wild-type and *Trif*$^{-/-}$ BMMs treated with viable *thyA*$^-$ *E. coli* for 1, 3 or 6 h (three biological replicates, numbered 1–3). A heat map of positive regulators/essential components (+) and negative regulators (−) of inflammasomes is shown. **h**, *Nlrp3* transcription at 1 h in BMMs. **i**, **j**, Serum levels of IL-6 and IL-1β 6 h after injection of 1×10^9 viable or 5×10^9 heat-killed *thyA*$^-$ *E. coli* (**i**), and splenic bacterial burdens 72 h after injection of 1×10^8 non-auxotroph *E. coli* (**j**) into wild-type, *Trif*$^{-/-}$, *Asc*$^{-/-}$ and *Nlrp3*$^{-/-}$ mice are shown. Each symbol represents one mouse. *, $P \leq 0.05$; **, $P \leq 0.01$; ***, $P \leq 0.001$. NS, not statistically significant. Hash symbol indicates not detected. All bars represent mean ± s.e.m.

Injection of viable and heat-killed *thyA*$^-$ *E. coli* into mice induced similarly high serum levels of IL-6 (Fig. 2i). In contrast, circulating IL-1β was detected only in mice infected with viable bacteria (Fig. 2i), whereas IFN-β levels were undetectable in all groups (data not shown). Confirming our results *in vitro*, production of IL-1β (but not IL-6) *in vivo* also required TRIF, ASC and NLRP3 (Fig. 2i). Injection of non-pathogenic *S. enterica* Typhimurium induced serum IL-1β levels comparable to those elicited by *thyA*$^-$ *E. coli*, which similarly depended on TRIF (Supplementary Fig. 12). Although pathogenic *S. enterica* Typhimurium elicited higher levels of serum IL-1β than non-pathogenic *Salmonella*, this response was also severely reduced in *Trif*$^{-/-}$ mice, suggesting a previously unappreciated role for TRIF in *Salmonella* infection (Supplementary Fig. 12). Importantly, deficiency in TRIF, ASC and NLRP3 impaired bacterial clearance during systemic infection with replication-sufficient non-pathogenic *E. coli* (Fig. 2j). This failure was more dramatic in *Trif*$^{-/-}$ than in *Asc*$^{-/-}$ or *Nlrp3*$^{-/-}$ mice, possibly due to the central upstream role of TRIF in inflammasome activation and IFN-β production.

The ability to sense microbial viability through pathways downstream of pattern recognition receptors indicates the existence of vita-PAMPs; that is, PAMPs associated with viable but not dead bacteria. In contrast to LPS and genomic DNA, which remained constant after killing *thyA*$^-$ *E. coli* with heat, total bacterial RNA was rapidly lost (Fig. 3a, b and Supplementary Fig. 13). Total RNA content was also lost with antibiotic treatment, and little ribosomal RNA (rRNA) remained after killing with ultraviolet irradiation and ethanol (Supplementary Fig. 14). Only fixation with paraformaldehyde (PFA) efficiently killed the bacteria (not shown) while preserving total RNA content (Supplementary Fig. 15a). Remarkably, unlike bacteria killed by other means, PFA-killed bacteria induced pyroptosis and IL-1β production to levels similar to those induced by viable bacteria (Supplementary Fig. 15b). Thus, the presence or absence of RNA correlated with the ability to activate pathways involved in sensing viability.

These results indicate that prokaryotic RNA represents a labile PAMP closely associated with bacterial viability that might signify microbial life to the immune system. Indeed, addition of purified total bacterial RNA fully restored the ability of heat-killed *thyA*$^-$ *E. coli* to induce pyroptosis, IL-1β and IFN-β production (Fig. 3c). These responses were dependent on TRIF, NLRP3 and caspase-1, just as those responses elicited by viable bacteria (Fig. 3d compared to Figs 1j and 2a–f). The NLRP3 inflammasome mediates recognition of viral RNA during influenza A infection[14]. Together with our results and those of others[15], this suggests a more general role for NLRP3 in responses to RNAs of microbial origin. RNA can activate the NLRP3 inflammasome when delivered into the cytosol (where NLRP3 is found) with transfection reagents[15]. In contrast, inflammasome activation by the combination of total bacterial RNA and dead *thyA*$^-$ *E. coli* did not require RNA transfection (Fig. 3c, d). Administration of total *E. coli* RNA alone or in combination with LPS (to mimic an *E.-coli*-derived PAMP plus RNA) had little effect on NLRP3 inflammasome activation unless the RNA was delivered to the cytosol using Lipofectamine (Supplementary Fig. 16) or in combination with ATP, as reported previously[15]. Thus, phagocytosis of viable bacteria is a natural context of bacterial-RNA-mediated NLRP3 inflammasome activation.

These findings raised the question as to how vita-PAMPs in phagolysosomes gain access to cytosolic receptors such as NLRP3 in the absence of invasion, auxiliary secretion systems or pore-forming toxins. To address this question, we exploited the pH-sensitive excitation spectrum of fluorescein: the acidic pH in phagolysosomes quenches fluorescence whereas release into the pH-neutral cytosol allows a regain in fluorescence[16]. Phagocytosis of avirulent *thyA*$^-$ *E. coli* in the presence of fluorescein-conjugated dextran (Fdx) consistently induced low-level release of Fdx into the cytosol of macrophages (Fig. 3e, f and Supplementary Fig. 17). This indicates that phagosomes carrying *E. coli* exhibit intrinsic leakiness, a property previously described for particles such as beads and crystals that induce phagolysosomal destabilization[16,17]. Interestingly, killed *E. coli* also induced Fdx release, although to a slightly lower extent than viable *E. coli* (Fig. 3e, f), demonstrating that phagosomal leakage occurs independently of bacterial viability. Therefore, RNA from viable

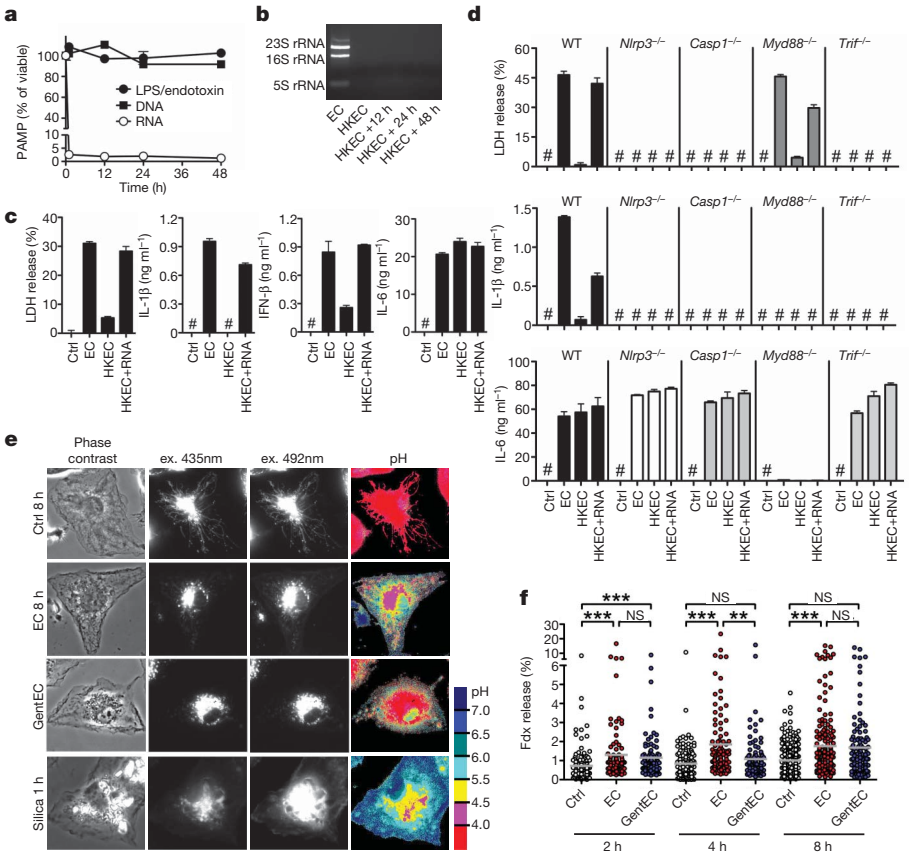


Figure 3 | Bacterial RNA is a vita-PAMP that accesses cytosolic receptors during phagocytosis and in the absence of virulence factors. **a**, LPS/endotoxin, genomic DNA and total RNA in *thyA*$^-$ *E. coli* before and at indicated times after heat killing. **b**, Agarose gel electrophoresis of *thyA*$^-$ *E. coli* total RNA before and after heat killing at 60 °C for 60 min followed by 4 °C incubation for the indicated times. **c**, **d**, LDH, IL-1β, IFN-β and IL-6 at 24 h in response to viable *thyA*$^-$ *E. coli* (EC), heat-killed *thyA*$^-$ *E. coli* (HKEC), or heat-killed *thyA*$^-$ *E. coli* with 10 μg ml^{-1} total RNA (HKEC+RNA). Hash symbol in **c** and **d** indicates not detected. Data in **a**–**d** are from murine BMMs and represent ≥5 experiments. **e**, Representative ratiometric epifluorescence imaging of murine BMMs at 8 h with Fdx alone (ctrl 8 h), Fdx and viable *thyA*$^-$ *E. coli* (EC 8 h) or gentamicin-killed *thyA*$^-$ *E. coli* (GentEC). Colour code indicates pH scale. Positive control is ground silica (silica 1 h). **f**, Quantification of cytosolic Fdx expressed as percentage of total Fdx per cell. Each dot represents the percentage of released Fdx per individual cell. Grey bars represent mean Fdx release. *, $P \leq 0.05$; **$P \leq 0.01$; ***, $P \leq 0.001$. All bars represent mean ± s.e.m.

bacteria could gain access to cytosolic receptors via intrinsic phagosomal leakage. These results may also explain the reported ability of phagosome-degraded mutants of *Listeria monocytogenes* or *Staphylococcus aureus* to induce a transcriptional response dependent on cytosolic NLRs[18,19].

Digestion of total RNA from *E. coli* with exonuclease RNase I and double-stranded RNA (dsRNA)-specific endonuclease RNase III abrogated LDH and IL-1β release, whereas DNase treatment had no effect (Fig. 4a). Of the *E. coli* RNA species, mRNA most potently induced pyroptosis as well as production of IL-1β and IFN-β. Small RNA (sRNA), or the most abundant RNA, ribosomal RNA (rRNA), had little or no detectable effects (Fig. 4b and Supplementary Fig. 18). *Escherichia coli* rRNA undergoes extensive modifications not found in mRNA[20], which may underlie the differential activity of these RNA species. The relative amount of mRNA was <1% of the total RNA and accordingly, mRNA was approximately 100-fold more effective than total RNA (Figs 3c and 4a, b and Supplementary Fig. 18).

In-vitro-transcribed single-stranded mRNA of the *E. coli* Gro operon (Supplementary Fig. 19a, b), which is strongly expressed upon phagocytosis of bacteria[21], induced caspase-1 cleavage and subsequent pyroptosis and IL-1β production when phagocytosed together with heat-killed *thyA*$^-$ *E. coli* (Fig. 4c, e and Supplementary Fig. 19c–e). The single-stranded Gro mRNA sequence had a predicted secondary structure with regions of high probability for base pairing (Fig. 4d), consistent with susceptibility of the stimulatory activity to RNase III treatment (Fig. 4a). Indeed, fully dsGro mRNA (Supplementary Fig. 19b) induced responses similar to single-stranded Gro mRNA of the appropriate length (Fig. 4e and Supplementary Fig. 19d). Other transcripts also induced such responses, showing that the immunostimulatory property is independent of RNA sequence (Fig. 4f).

Notably, eukaryotic RNA was unable to elicit the responses induced by *E. coli* mRNA (Fig. 4b). Unlike eukaryotic mRNA, triphosphate moieties at the 5′ end of bacterial mRNAs are not capped with 7-methyl-guanosine ($7m^7G$)[22], and might betray the prokaryotic origin of these transcripts[23]. However, neither treatment with calf intestinal phosphatase (CIP) nor capping affected the activity of Gro mRNA during phagocytosis of heat-killed *thyA*$^-$ *E. coli* (Fig. 4g). The stimulatory activity of purified *E. coli* total RNA or mRNA was also unaltered by CIP treatment (Supplementary Fig. 20a, b), arguing against a role for the RNA helicase retinoic acid inducible gene-I (RIG-I), which can induce interferon and IL-1β production but requires 5′-triphosphates for activation (Supplementary Fig. 20c)[23]. Moreover, TRIF and NLRP3 are dispensable for RIG-I function but are required for the stimulatory activity of bacterial RNA (Figs 2a, b and 3d). Interestingly, RNA can induce RIG-I-dependent IFN-β during infection with an invasive intracellular bacterium[24], indicating that the nature of microbial pathogenesis and the cellular context in which bacterial RNA is recognized may determine the choice of innate sensors engaged. In contrast to 5′-triphosphate removal, adding polyadenylyl groups to the 3′ end of Gro mRNA or purified *E. coli* mRNA abrogated IL-1β secretion and pyroptosis (Fig. 4g and Supplementary Fig. 21). Thus, absence of 3′-polyadenylation[22] may allow specific detection of prokaryotic mRNA during infection. Additional features may distinguish self from microbial RNAs such as internal naturally occurring nucleoside modifications in eukaryotic RNA[25–27].

To test the impact of vita-PAMPs on adaptive immunity, we immunized mice with either viable or dead *thyA*$^-$ *E. coli*, or a combination of dead *thyA*$^-$ *E. coli* and purified total bacterial RNA (Supplementary Fig. 22). Whereas all three vaccines induced similar polyclonal anti-*E. coli* IgM responses, production of class-switched IgG subclasses was strongly enhanced in response to vaccination with viable compared to killed *E. coli* (Fig. 4h). Adding total bacterial RNA to killed *thyA*$^-$ *E. coli* elevated IgG1, IgG2c, IgG2b and IgG3 antibody titres to or above the levels in mice immunized with viable *thyA*$^-$ *E. coli*. Thus, innate detection of bacterial viability leads to robust activation of a humoral adaptive response. These findings indicate that bacterial RNA can augment killed vaccines to perform as well as live ones.

Our findings reveal an inherent ability of the immune system to distinguish viable from dead microorganisms. The presence of live bacteria in sterile tissues, regardless of whether these (still) express

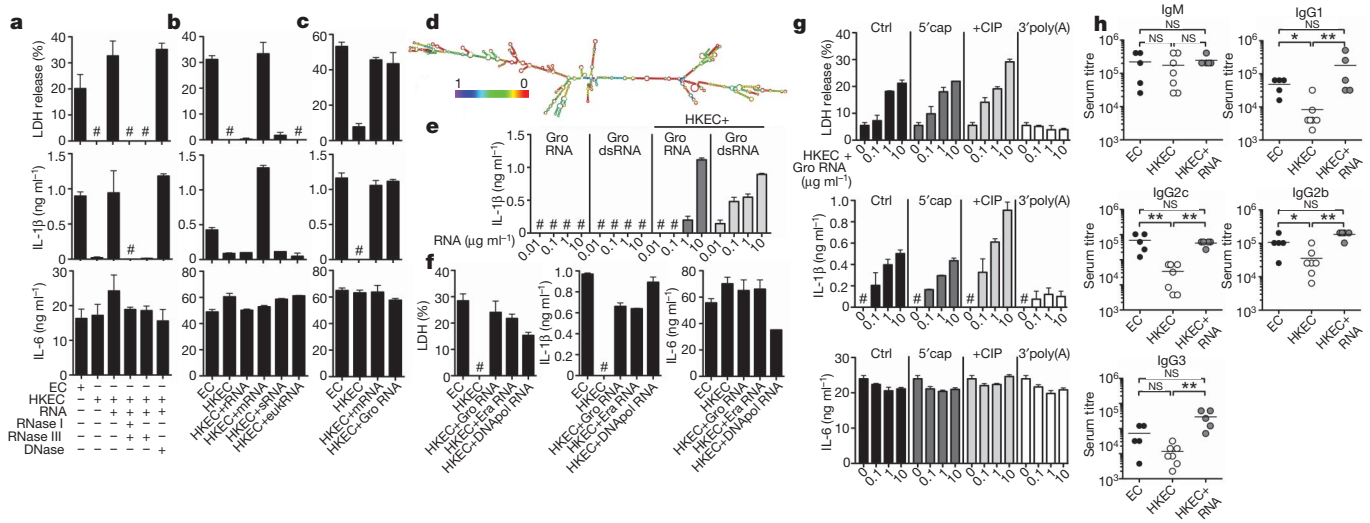


Figure 4 | Bacterial mRNA constitutes an active vita-PAMP. **a**–**c**, **e**–**g**, LDH, IL-1β and IL-6 at 24 h. **a**, Total *thyA*$^-$ *E. coli* RNA treated with RNase I and RNase III, RNase III alone, or DNase before stimulation of BMDCs. **b**, BMMs treated with viable or heat-killed *thyA*$^-$ *E. coli*, or heat-killed *thyA*$^-$ *E. coli* with 0.1 μg ml^{-1} of different bacterial RNA (ribosomal RNA (rRNA), mRNA, small RNA (sRNA) or eukaryotic RNA (eukRNA)). **c**, BMDC responses. Gro RNA indicates *in-vitro*-transcribed *E. coli* Gro operon RNA. **d**, Predicted secondary structure of Gro RNA. The colour code indicates base pairing probability. **e**, BMMs treated with *in-vitro*-transcribed Gro RNA or Gro dsRNA alone or with heat-killed *thyA*$^-$ *E. coli*. **f**, BMDC responses. Era RNA and DNApol RNA indicate *in-vitro*-transcribed *E. coli* Era GTPase and DNA polymerase III RNA, respectively. **g**, BMMs treated with different doses of unmodified (ctrl, control) or modified Gro RNA with heat-killed *thyA*$^-$ *E. coli* (5′ cap, 5′ m^7G capping; CIP, calf intestinal phosphatase; 3′ poly(A), 3′-polyadenylation). For **a**–**g**, the hash symbol indicates not detected; all RNA at 10 μg ml^{-1} except as noted; data represent ≥5 experiments. **h**, Mice vaccinated and boosted twice with viable *thyA*$^-$ *E. coli* (EC), heat-killed *thyA*$^-$ *E. coli* (HKEC) or heat-killed *thyA*$^-$ *E. coli* with 30 μg total purified bacterial RNA (HKEC+RNA) (vaccination regimen is given in Supplementary Fig. 22). Class-specific anti-*E. coli* antibody serum titres at 25 days are shown. *, $P \leq 0.05$; **, $P \leq 0.01$; ***, $P \leq 0.001$. All bars represent mean ± s.e.m.

virulence factors, poses an acute threat that must be dealt with by an aggressive immune response. Dead bacteria, on the other hand, would signify a successful immune response that can now subside. Detection of vita-PAMPs within sterile tissues signifies microbial viability. Other vita-PAMPs may exist in the form of second messengers like cyclic di-adenosine or di-guanosine monophosphates[7,28] or quorum-sensing molecules[7]. The extent to which vita-PAMPs contribute to the host response during natural infection with pathogenic bacteria, relative to other stimuli such as the activity of virulence factors, is an important issue that requires further investigation. Given that bacteria tightly regulate their virulence via multiple mechanisms in response to different environmental signals and inside a host organism during infection[29,30], detection of invariant vita-PAMPs essential to bacterial survival may be a non-redundant fail-safe strategy for host protection.

METHODS SUMMARY

Cells were infected with *E. coli* DH5α *thyA*$^-$ at a multiplicity of infection of 20 for 24 h unless stated otherwise. Supernatants were assayed for cytokines by ELISA. Genome-wide transcriptional analysis of murine bone-marrow-derived macrophages (BMMs) at 0, 1, 3 and 6 h after infection was carried out on Affymetrix GeneChip Mouse Gene 1.1 ST 24-array plates. Phagosomal leakage in BMMs was detected by measuring Fdx release using a modified method previously described[16]. In brief, BMMs were treated with *thyA*$^-$ *E. coli* in the presence of 0.167 mg ml^{-1} Fdx and imaged with excitation at 440 nm (pH insensitive) and 485 nm (pH sensitive). Fluorescence intensity ratios at 485 nm/440 nm were converted into pH maps and the percentage of Fdx release calculated (total intensity of pixels containing released Fdx/total Fdx intensity). Bacterial RNA was extracted from *E. coli* using the e.z.n.a RNA kit (Omega) and *in vitro* transcription of bacterial genes carried out using the MEGAscript kit (Ambion) followed by DNase digestion and RNA purification using the MEGAclear kit (Ambion). RNA polyadenylation was performed with the poly(A)-tailing kit (Ambion). Vaccinations were performed as a prime–boost regimen (see Methods). C57BL/6J and *P2rx7*$^{-/-}$ mice were purchased from the Jackson Laboratory. *Myd88*$^{-/-}$ and *Trif*$^{-/-}$ mice were provided by S. Akira, *Trif*$^{-/-}$×*Myd88*$^{-/-}$ by R. Medzhitov, *Nlrp3*$^{-/-}$, *Asc*$^{-/-}$ and *Nlrc4*$^{-/-}$ by Millenium, and *Casp1*$^{-/-}$ by R. Flavell. Animal care and experimentation were performed in accordance with approved MSSM Institutional Animal Care and Use Committee protocols.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 30 April 2010; accepted 24 March 2011.
Published online 22 May 2011.

1. Brockstedt, D. G. *et al.* Killed but metabolically active microbes: a new vaccine paradigm for eliciting effector T-cell responses and protective immunity. *Nature Med.* **11,** 853–860 (2005).
2. Cheers, C. & Zhan, Y. How do macrophages distinguish the living from the dead? *Trends Microbiol.* **4,** 453–455 (1996).
3. Detmer, A. & Glenting, J. Live bacterial vaccines—a review and identification of potential hazards. *Microb. Cell Fact.* **5,** 23 (2006).
4. Kawamura, I. *et al.* Antigen provoking gamma interferon production in response to *Mycobacterium bovis* BCG and functional difference in T-cell responses to this antigen between viable and killed BCG-immunized mice. *Infect. Immun.* **62,** 4396–4403 (1994).
5. Lauvau, G. *et al.* Priming of memory but not effector CD8 T cells by a killed bacterial vaccine. *Science* **294,** 1735–1739 (2001).
6. von Koenig, C. H., Finger, H. & Hof, H. Failure of killed *Listeria monocytogenes* vaccine to produce protective immunity. *Nature* **297,** 233–234 (1982).
7. Vance, R. E., Isberg, R. R. & Portnoy, D. A. Patterns of pathogenesis: discrimination of pathogenic and nonpathogenic microbes by the innate immune system. *Cell Host Microbe* **6,** 10–21 (2009).
8. Medzhitov, R. Approaching the asymptote: 20 years later. *Immunity* **30,** 766–775 (2009).
9. Takeuchi, O. & Akira, S. Pattern recognition receptors and inflammation. *Cell* **140,** 805–820 (2010).
10. Mariathasan, S. & Monack, D. M. Inflammasome adaptors and sensors: intracellular regulators of infection and inflammation. *Nature Rev. Immunol.* **7,** 31–40 (2007).
11. Schroder, K. & Tschopp, J. The inflammasomes. *Cell* **140,** 821–832 (2010).
12. Wing, H. J., Yan, A. W., Goldman, S. R. & Goldberg, M. B. Regulation of IcsP, the outer membrane protease of the *Shigella* actin tail assembly protein IcsA, by virulence plasmid regulators VirF and VirB. *J. Bacteriol.* **186,** 699–705 (2004).
13. Zhou, R., Yazdi, A. S., Menu, P. & Tschopp, J. A role for mitochondria in NLRP3 inflammasome activation. *Nature* **469,** 221–225 (2011).
14. Pang, I. K. & Iwasaki, A. Inflammasomes as mediators of immunity against influenza virus. *Trends Immunol.* **32,** 34–41 (2011).
15. Kanneganti, T. D. *et al.* Bacterial RNA and small antiviral compounds activate caspase-1 through cryopyrin/Nalp3. *Nature* **440,** 233–236 (2006).
16. Davis, M. J. & Swanson, J. A. Technical advance: Caspase-1 activation and IL-1β release correlate with the degree of lysosome damage, as illustrated by a novel imaging method to quantify phagolysosome damage. *J. Leukoc. Biol.* **88,** 813–822 (2010).
17. Hornung, V. *et al.* Silica crystals and aluminum salts activate the NALP3 inflammasome through phagosomal destabilization. *Nature Immunol.* **9,** 847–856 (2008).
18. Herskovits, A. A., Auerbuch, V. & Portnoy, D. A. Bacterial ligands generated in a phagosome are targets of the cytosolic innate immune system. *PLoS Pathog.* **3,** e51 (2007).
19. Shimada, T. *et al. Staphylococcus aureus* evades lysozyme-based peptidoglycan digestion that links phagocytosis, inflammasome activation, and IL-1β secretion. *Cell Host Microbe* **7,** 38–49 (2010).
20. Piekna-Przybylska, D., Decatur, W. A. & Fournier, M. J. The 3D rRNA modification maps database: with interactive tools for ribosome analysis. *Nucleic Acids Res.* **36,** D178–D183 (2008).
21. Buchmeier, N. A. & Heffron, F. Induction of *Salmonella* stress proteins upon infection of macrophages. *Science* **248,** 730–732 (1990).
22. Belasco, J. G. All things must pass: contrasts and commonalities in eukaryotic and bacterial mRNA decay. *Nature Rev. Mol. Cell Biol.* **11,** 467–478 (2010).
23. Rehwinkel, J. & Reis e Sousa, C. RIGorous detection: exposing virus through RNA sensing. *Science* **327,** 284–286 (2010).
24. Monroe, K. M., McWhirter, S. M. & Vance, R. E. Identification of host cytosolic sensors and bacterial factors regulating the type I interferon response to *Legionella pneumophila*. *PLoS Pathog.* **5,** e1000665 (2009).
25. Nallagatla, S. R., Toroney, R. & Bevilacqua, P. C. A brilliant disguise for self RNA: 5′-end and internal modifications of primary transcripts suppress elements of innate immunity. *RNA Biol.* **5,** 140–144 (2008).
26. Anderson, B. R. *et al.* Incorporation of pseudouridine into mRNA enhances translation by diminishing PKR activation. *Nucleic Acids Res.* **38,** 5884–5892 (2010).
27. Kariko, K., Buckstein, M., Ni, H. & Weissman, D. Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA. *Immunity* **23,** 165–175 (2005).
28. Woodward, J. J., Iavarone, A. T. & Portnoy, D. A. c-di-AMP secreted by intracellular *Listeria monocytogenes* activates a host type I interferon response. *Science* **328,** 1703–1705 (2010).
29. Gripenland, J. *et al.* RNAs: regulators of bacterial virulence. *Nature Rev. Microbiol.* **8,** 857–866 (2010).
30. Raskin, D. M., Seshadri, R., Pukatzki, S. U. & Mekalanos, J. J. Bacterial genomics and pathogen evolution. *Cell* **124,** 703–714 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to R. Medzhitov and J. C. Kagan for critical reading of the manuscript; C. B. Lopez for *Irf3*$^{-/-}$ mice; D. M. Monack for *Salmonella* Δ*Spi1*Δ*Spi2*; M. B. Goldberg for *Shigella* BS103; and D. A. Portnoy for *Listeria* Δ*Hly*Δ*fliC*. We thank M. Rivieccio, I. Brodsky, M. Blander, S. J. Blander, J. Sander and Blander laboratory members for insightful discussions, help and support. L.E.S. was supported by Deutsche Forschungsgemeinschaft grant SA-1940/1-1, D.A. by fellowships from the Academic Medical Center and the Landsteiner Foundation for Blood Research, and M.V.B. and M.M. by the Netherlands Nutrigenomics Centre. This work was supported by NIH grant AI080959A and the Kinship Foundation Searle Scholar award to J.M.B.

Author Contributions L.E.S. and J.M.B. designed experiments and directed the study. L.E.S. performed all experiments. M.J.D. and L.E.S. performed experiments measuring lysosomal leakage. J.A.S. helped with the design and analysis of the lysosomal leakage experiments. M.V.B. performed gene microarray analysis. M.V.B. and M.M. analysed the gene microarray data and helped with data interpretation. D.A. and J.M.B. performed experiments during the development phase of the project, and C.C.D. helped with the design of RNA-related experiments. B.R. provided bone marrow progenitor cells from *Nlrp3*$^{-/-}$, *Asc*$^{-/-}$ and *Casp1*$^{-/-}$ mice. L.E.S., D.A. and J.M.B. wrote the manuscript. J.M.B. conceived of the study.

Author Information Affymetrix Microarray data have been deposited with the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE27960. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.M.B. (julie.blander@mssm.edu).

METHODS

Cells. Bone-marrow-derived dendritic cell (BMDC) cultures were grown as previously described[31] in RPMI 1640 supplemented with granulocyte-macrophage colony-stimulating factor (GM-CSF) and 5% fetal bovine serum (FBS), plus 100 μg ml^{-1} penicillin, 100 μg ml^{-1} streptomycin, 2 mM L-glutamine, 10 mM HEPES, 1 nM sodium pyruvate, 1% MEM non-essential amino acids, and 2.5 μM β-mercaptoethanol (all Sigma). Semi-adherent cells were harvested on ice on day 5 and re-plated immediately in fresh RPMI 1640 medium containing 10% FBS at 5×10^5 cells per well in 24-well tissue-culture-treated plates. Stimuli were added immediately after re-plating in the same medium and the cells were centrifuged for 2 min at 2,000 r.p.m. Murine macrophages were derived from the bone marrow (BMMs) of C57BL/6J, *Myd88*$^{-/-}$, *Trif*$^{-/-}$, *Trif*$^{-/-}$×*Myd88*$^{-/-}$, *Nlrp3*$^{-/-}$, *Asc*$^{-/-}$ or *Casp1*$^{-/-}$ mice, as described previously[32], in RPMI 1640 supplemented with M-CSF and 10% FBS, plus 100 μg ml^{-1} penicillin, 100 μg ml^{-1} streptomycin, 10 mM HEPES and 1 nM sodium pyruvate (all Sigma). For some experiments macrophages were derived from the bone marrow of *Irf3*$^{-/-}$ or *P2rx7*$^{-/-}$ mice. Peritoneal macrophages were harvested 72 h after intraperitoneal injection of 1 ml thioglycollate (BD Bioscience), grown overnight in RPMI 1640 medium supplemented with 10% FBS and 100 μg ml^{-1} penicillin and 100 μg ml^{-1} streptomycin, hereafter referred to as 'complete medium'. Mouse embryonic fibroblasts (MEFs) deficient for RIG-I (*RIG-I*$^{-/-}$) were provided by A. Ting with permission from S. Akira, and grown in DMEM medium containing 10% FBS and 100 μg ml^{-1} penicillin, 100 μg ml^{-1} streptomycin.

Mice. C57BL/6J and *P2rx7*$^{-/-}$ mice were purchased from Jackson Laboratories. *Myd88*$^{-/-}$ and *Trif*$^{-/-}$ mice were originally provided by S. Akira; *Myd88*$^{-/-}$ and *Trif*$^{-/-}$ mice were interbred to homozygosity to generate *Trif*$^{-/-}$×*Myd88*$^{-/-}$ mice, and were provided by R. Medzhitov. *Nlrp3*$^{-/-}$, *Asc*$^{-/-}$ or *Casp1*$^{-/-}$ bone marrow was provided by B. Ryffel and mice for *in vivo* studies were acquired from R. Flavell (through Millenium) and have been described previously[33,34]. *Irf3*$^{-/-}$ mice were provided by C. B. Lopez and were previously described[35]. We used 8–10-week-old animals for all experiments. All experiments were approved by the institutional ethics committee and carried out in agreement with the 'Guide for the Care and Use of Laboratory Animals' (NIH publication 86-23, revised 1985).

Bacteria. *Escherichia coli* K12, strain DH5α were purchased from Invitrogen. Naturally occurring thymidine auxotrophs (*thyA*$^-$) were selected on Luria-Bertani (LB) agar plates containing 50 μg ml^{-1} trimethoprim and 500 μg ml^{-1} thymidine (both Sigma). Auxotrophy was confirmed by inoculation and overnight culture of single colonies in LB medium. *thyA*$^-$ *E. coli* grew only in the presence of thymidine and were resistant to trimethoprim. For phagocytosis experiments, *thyA*$^-$ *E. coli* were grown to mid-log phase, washed three times in phosphate buffered saline (PBS) to remove thymidine and LB salts before addition to cells. For heat killing, *thyA*$^-$ *E. coli* were grown to log phase, washed and re-suspended in PBS at an optical density at 600 nm (OD_{600}) of 0.6, and subsequently incubated at 60 °C for 60 min. *thyA*$^-$ heat-killed *E. coli* were stored up to 18 h at 4 °C or used immediately after cooling. Efficient killing was confirmed by overnight plating on thymidine/trimethoprim-supplemented LB-agar plates. For gentamicin killing, *thyA*$^-$ *E. coli* were grown to mid-log phase, washed and re-suspended in LB medium containing thymidine, trimethoprim and 50 μg ml^{-1} gentamicin sulphate and incubated in a shaking incubator at 37 °C overnight. Ethanol killing was carried out by re-suspending log phase *thyA*$^-$ *E. coli* in 70% ethanol for 10 min, followed by extensive washing in PBS. For ultraviolet killing, log phase *thyA*$^-$ *E. coli* were re-suspended in PBS at an OD_{600} of 0.6, ultraviolet-irradiated with 1,000 mJ cm^{-2} in a Petri dish followed by washing with PBS. Paraformaldehyde (PFA) fixation was performed by re-suspending log-phase *thyA*$^-$ *E. coli* in 4% PFA in PBS for 10 min followed by extensive washing and re-suspension in PBS. *Shigella flexneri* virulence plasmid-cured strain BS103 was provided by M. B. Goldberg[12,36]. *thyA*$^-$ *S. flexneri* were selected similarly to *thyA*$^-$ *E. coli*. D. M. Monack provided *Salmonella enterica* serovar Typhimurium, strain SL1344 Δ*Spi1*Δ*Spi2*, lacking the *Salmonella* pathogenicity island SPI-1 and SPI-2 type-III secretion systems[37]. SL1344 Δ*Spi1*Δ*Spi2* was grown in LB medium containing 25 μg ml^{-1} kanamycin and 12 μg ml^{-1} tetracycline. *Listeria monocytogenes* Δ*Hly*Δ*fliC* lacking listeriolysin O (LLO) and flagellin expression were provided by D. Portnoy[38].

Treatment of macrophages and dendritic cells with viable and killed bacteria. Macrophages were detached and re-plated 4 h before the experiment. BMDCs were re-plated immediately before addition of bacteria or soluble ligands. Unless stated otherwise, bacteria were used at a multiplicity of infection of 20. All experiments were carried out in antibiotic-free 'complete medium'. One hour after addition of bacteria, penicillin (100 μg ml^{-1}) and streptomycin (100 μg ml^{-1}) were added to the medium to kill any remaining extracellular bacteria. Alternatively, gentamicin sulphate (50 μg ml^{-1}) was used. We also compared this approach to washing the cells and replacing the antibiotic-free medium with penicillin/streptomycin containing medium after 1 h and found no differences with regards to the cellular responses measured. Supernatants were collected 24 h after the addition of the bacteria unless stated otherwise in the figure legends.

Cytokine enzyme-linked immunosorbent assays. Supernatants from cultured BMMs or BMDCs were collected at 24 h after stimulation or at the times indicated. Enzyme-linked immunosorbent assay (ELISA) antibody pairs used for IL-6, IL-1β and TNF-α were as listed below. All ELISA antibodies were used at 2 μg ml^{-1} capture and 0.5 μg ml^{-1} detection, with the exception of IL-6 capture, which was used at 1 μg ml^{-1}. Detection antibodies were biotinylated and labelled by streptavidin-conjugated horseradish peroxidase (HRP), and visualized by the addition of *o*-phenylenediamine dihydrochloride (Sigma) (from tablets) or 3,3′, 5,5′-tetramethylbenzidine solution (TMB, KPL). Colour development was stopped with 3 M H_2SO_4 or TMB-Stop Solution (KPL), respectively. Recombinant cytokines served as standards and were purchased from Peprotech. Absorbances at 492 or 450 nm were measured, respectively, on a tunable microplate reader (VersaMax, Molecular Devices). Cytokine supernatant concentrations were calculated by extrapolating absorbance values from standard curves where known concentrations were plotted against absorbance using SoftMax Pro 5 software. Capture/detection antibody pairs were as follows. IL-6, MP5-20F3/MP5-32C11 (BDPharmingen); IL-1β, B12/rabbit polyclonal antibody (eBioscience); TNF-α, TN3-19/rabbit polyclonal antibody (eBioscience). IFN-β production was measured from supernatants using the VeriKine Mouse IFN-Beta ELISA Kit (PBL Interferon source) following manufacturer's instructions.

Anti-*E. coli* antibody ELISA. 96-well microtitre plates were coated overnight with *E. coli* lysates (3 μg ml^{-1}) that we generated from log-phase cultures of *thyA*$^-$ *E. coli*. Serum samples from immunized mice were serially diluted (12 dilutions) and incubated in the pre-coated plates for 12 h at 4 °C followed by washing and incubation with rabbit anti-mouse isotype-specific Ig–HRP (Southern Biotech) for 1 h. Bound rabbit anti-mouse Ig–HRP was visualized by the addition of *o*-phenylenediamine dihydrochloride (Sigma) from tablets, and the anti-*E. coli* antibody titres for each mouse were determined by absorbance readings at 490 nm.

Measurement of inflammatory cell death. Cell death of macrophages or BMDCs was measured using the Cytotox96 cytotoxicity assay (Promega) following manufacturer's instructions. The assay measures the release of lactate dehydrogenase (LDH) into the supernatant calculated as the percentage of total LDH content, measured from cellular lysates (100%). LDH released by unstimulated cells was used for background correction.

Flow cytometric assessment of cell death. Cells were stimulated overnight, stained for Annexin V/7AAD using the Annexin V-PE/7AAD Apoptosis Detection kit (BD Pharmingen), and analysed by flow cytometry (FACSCalibur, BD).

Flow cytometric measurement of ROS production. BMMs were loaded with the ROS indicator dye H2DCFDA (Molecular Probes/Invitrogen, 10 mM in PBS) for 30 min followed by a recovery time of 30 min in fresh pre-warmed 'complete medium'. BMMs were then stimulated with viable or heat killed *E. coli* for 60 min, washed and analysed by flow cytometry (FACSCalibur, BD).

Western blots. For detection of caspase-1, protein extracts were separated on 4–12% SDS-gradient gels (Invitrogen). For detection of all other proteins, samples were run on 10% SDS-polyacrylamide gels. Proteins were transferred to PVDF membranes (Millipore). Membranes were blocked with 5% milk in PBS and probed with the following antibodies: caspase-1 p10 (M-20)/rabbit polyclonal antibody, IκBα (C-21)/rabbit polyclonal antibody (both from Santa Cruz Biotechnologies), phospho-IRF3 (Ser 396)/rabbit polyclonal antibody, IRF3/rabbit polyclonal antibody, phospho-p38 MAPK (Thr 180/Tyr 182)/rabbit polyclonal antibody, p38 MAPK/rabbit polyclonal antibody (all from Cell Signalling Technology), α-tubulin (DM1A)/rabbit monoclonal antibody (Novus Biologicals).

Real-time PCR. Total RNA was isolated from macrophages using the RNeasy kit (Qiagen). Contaminating genomic DNA was removed by DNase digestion (DNase I, Promega). Reverse transcription was performed using Superscript III (Invitrogen) and cDNA was used for subsequent real-time PCR reactions. Quantitative real-time RT–PCR was conducted on an ABI Prism 7900 instrument using the Maxima SYBR green qPCR Master Mix (Fermentas) with the following primer pairs. β-Actin, FW 5′-GAAGTCCCTCACCCTCCCAA-3′, RV 5′-GGCATGGACGCGACCA-3′; *Il1b*, FW 5′-AAAGACGGCACACCCACCCTGC-3′, RV 5′-TGTCCTGACCACTGTTGTTTCCCAG-3′; *Ifnb*, FW 5′-GCACTGGGTGGAAT-3′, RV 5′-TTCTGAGGCATCAA-3′; *Nlrp3*, FW 5′-CGAGACCTCTGGGAAAAAGCT-3′, RV 5′-GCATACCATAGAGGAATGTGATGTACA-3′. All reactions were performed in duplicates and the samples were normalized to β-actin. 'Fold inductions' were calculated using the $\Delta\Delta C^t$ method relative to unstimulated BMMs.

Transcriptome analysis. BMMs derived from wild-type or *Trif*$^{-/-}$ mice were stimulated with viable *E. coli* for 0, 1, 3 or 6 h and total RNA was extracted using the RNeasy kit (Qiagen). RNA from three independent experiments was used for

transcriptional analysis. RNA integrity was checked on an Agilent 2100 Bioanalyser (Agilent Technologies) with 6000 Nano Chips. RNA was judged as suitable only if samples showed intact bands of 18S and 28S ribosomal RNA subunits, displayed no chromosomal peaks or RNA degradation products, and had a RNA integrity number (RIN) above 8.0.

One-hundred nanograms of RNA were used for whole-transcript cDNA synthesis with the Ambion WT expression kit (Applied Biosystems). Hybridization, washing and scanning of an Affymetrix GeneChip Mouse Gene 1.1 ST 24-array plate was carried out according to standard Affymetrix protocols on a GeneTitan instrument (Affymetrix).

Packages from the Bioconductor project, integrated in an in-house developed management and analysis database for microarray experiments, were used for analysis of the scanned arrays[39]. Arrays were normalized using the Robust Multi-array Average method[40,41]. Probe sets were defined according to ref. 42. With this method probes are assigned to unique gene identifiers, in this case Entrez IDs. The probes on the Gene 1.1 ST arrays represent 19,807 genes that have at least 10 probes per identifier. For the analysis, only genes that had an intensity value of >20 on at least two arrays were taken into account. In addition, the interquartile range of $\log_2$ intensities had to be at least 0.25. These criteria were met by 9,921 genes. Changes in gene expression are represented as signal log ratios between treatment and control. Multiple Experiment Viewer software (MeV 4.6.1) was used to create heatmaps[43,44]. Genes were clustered by average linkage hierarchical clustering using Pearson correlation. Significantly regulated genes were identified by intensity-based moderated *t*-statistics[45]. Obtained *P*-values were corrected for multiple testing by a false discovery rate method[46].

IFN-regulated genes were identified using the Interferome database (http://www.interferome.org)[47] and grouped in a heat map. Rel/NF-κB target genes were identified using another online database (http://bioinfo.lifl.fr/NF-KB/) which compiles Rel/NF-κB target genes identified by various groups[48] (http://people.bu.edu/gilmore/nf-kb/index.html). Inflammasome-related genes were compiled based on the current literature[11,49].

Measuring release from bacterial phagosomes. Measurement of fluorescein–dextran (Fdx) release from macrophage phagosomes was performed using a modified method described previously[16]. BMMs were plated onto Mat-tek coverslip dishes (MatTek Corp.) and incubated overnight. BMMs were stimulated with viable or gentamicin-killed red fluorescent protein (RFP)-expressing *thyA*$^-$ *E. coli* in the presence of 0.167 mg ml^{-1} Fdx in 200 µl of medium. After 120 min of co-culture, additional Fdx and gentamicin containing medium was added to the coverslip dishes to prevent drying and to prevent bacterial overgrowth. Cells were imaged after 2, 4 and 8 h to measure release of Fdx. Microscopic imaging was performed on an IX70 inverted microscope (Olympus) equipped with an X-cite 120 metal halide light source (EXFO) and excitation and emission filter wheels. Phase contrast and two fluorescent images were acquired for each field of cells. The fluorescent images used the same emission settings, but used different excitation band-pass filters. Fdx fluorescence intensity using an excitation filter centred at 440 nm is relatively insensitive to pH, whereas fluorescence intensity using an excitation filter centred at 485 nm is very sensitive to pH. The ratio of fluorescence intensity at 485 nm divided by 440 nm was converted to into pH maps using calibration curves generated by imaging BMMs with Fdx-containing compartments at a series of fixed pH conditions. As described previously[16], pixels with pH above 5.5 were designated as representing Fdx which has been released from endolysosomal compartments. The percentage of Fdx release was calculated by dividing the total intensity of pixels containing released Fdx by the total Fdx intensity for each cell.

Infections and vaccinations. For measurements of systemic cytokine levels, C57BL/6J wild-type, *Trif*$^{-/-}$, *Asc*$^{-/-}$ or *Nlrp3*$^{-/-}$ mice were injected with 1×10^9 viable or 5×10^9 heat-killed *thyA*$^-$ *E. coli*, respectively. Blood samples were drawn 6 h after infection, and cytokine concentrations were measured by ELISA. For determination of bacterial clearance, we infected mice with 1×10^8 viable replication-sufficient *E. coli* by intraperitoneal injection. Mice were monitored daily and moribund animals were killed according to humane criteria established and approved by our institutional IACUC committee. After 60 h, animals were killed and the spleens were explanted, homogenized, serially diluted and plated on LB-agar plates overnight followed by colony forming units (c.f.u.) counting.

For vaccinations, we followed a prime–boost regimen as shown in the schematic in Fig. 4h that was adopted from a previous study[50]. In brief, mice received an initial vaccination intraperitoneally with 5×10^7 c.f.u. of viable or heat-killed *thyA*$^-$ *E. coli* or a combination of 5×10^7 c.f.u. heat-killed *thyA*$^-$ *E. coli* and 30 µg of purified *E. coli* total RNA, followed by two boosts (5×10^6 c.f.u.) after 10 and 20 days. Polyclonal class-specific anti-*E. coli* antibody production was measured in the serum after 25 days by ELISA.

Bacterial RNA. Total bacterial RNA was isolated from *thyA*$^-$ *E. coli* using the e.z.n.a. Bacterial RNA Kit (Omega Bio-Tek), following the manufacturer's instructions. Contaminating DNA was removed by DNase digestion (TURBO DNase, Ambion/Applied Biosystems). Alternatively, total purified *E. coli* (DH5α) RNA was purchased from Ambion/Applied Biosystems, and similar results were obtained. Fractionation of bacterial RNA species was performed as follows. First, ribosomal 16S and 23S RNA (rRNA) was removed by a magnetic bead-based capture hybridization approach using the MICROBExpress kit (Ambion/Applied Biosystems). The enriched RNA was then separated into messenger RNA (mRNA) and small RNA (sRNA, including 5S rRNA) using the MEGAClear kit (Ambion/Applied Biosystems). All separated RNA fractions were precipitated with ammonium acetate and re-suspended in nuclease-free water. RNA concentration and purity were determined by measuring the absorbance at 260/280 and 260/230 nm. RNA preparations were further visualized by 1% agarose gel electrophoresis.

***In vitro* RNA transcription.** The *E. coli* Gro operon encoding the bacterial chaperonins GroEL and GroES, the GTPase Era operon or the DNA polymerase III operon were PCR amplified from genomic DNA isolated from *thyA*$^-$ *E. coli* using primer pairs containing a T7 promoter sequence (T7) in either the FW or both FW and RV primer. Gro-FWT7 5′-TAATACGACTCACTATAGGGCACCAGCCGGGAAACCACG-3′; Gro-RVT7 5′-TAATACGACTCACTATAGGAAAAGAAAAACCCCCAGACAT-3′; Gro-RV 5′-AGATGACCAAAAGAAAAACCCCCAGACATT-3′; Era-FWT7 5′-TAATACGACTCACTATAGGGCATATGAGCATCGATAAAAGTTAC-3′; Era-RV 5′-TTTAAAGATCGTCAACGTAACCGAG-3′; DNApol-FWT7 5′-TAATACGACTCACTATAGGGATGTCTGAACCACGTTTCGT-3′; DNApol-RV 5′-AGTCAAACTCCAGTTCCACCTGCTCCGAA-3′.

PCR fragments were purified using the Nucleospin Extract II PCR purification kit (Macherey-Nagel), and used as DNA templates for *in vitro* transcription. *In vitro* transcription was performed using the MEGAscript kit T7 (Ambion/Applied Biosystems) following the manufacturer's instructions. DNA templates generated with Gro-FWT7 and Gro-RV primers only contained a T7 promoter site in the sense strand and yielded single-stranded RNA, whereas PCR templates generated with Gro-FWT7 and Gro-RVT7 primers contained T7 promoter sequences in both strands, allowing transcription of two complementary strands, yielding double-stranded RNA. For generation of 5′-capped RNA, m7G(5′)ppp(5′)G cap analogue (Ambion/Applied Biosystems) was included in the transcription reaction at a GTP:cap ratio of 1:4.

RNA digestion, dephosphorylation and polyadenylation. *In-vitro*-transcribed Gro RNA, total *E. coli* RNA or *E. coli* mRNA were digested using RNase I (Promega) and RNase III (Ambion/Applied Biosystems). To remove 5′-triphosphates, RNA dephosphorylation was performed by incubating 10 µg *in-vitro*-transcribed RNA or total *E. coli* RNA or 1 mg of *E. coli* mRNA with 30 U of calf intestinal alkaline phosphatase (CIP, New England Biolabs) for 2 h at 37 °C, as described previously[51]. Polyadenylation of *in-vitro*-transcribed and purified bacterial mRNA was performed using the poly(A) Tailing kit (Ambion) following the manufacturer's instructions.

Transfection of macrophages and MEFs. For direct cytosolic delivery of total purified *E. coli* RNA or *in-vitro*-transcribed Gro RNA, 5×10^5 BMMs or 2×10^5 MEFs were transfected with 1 mg of RNA using 2 µl of Lipofectamine 2000 (Invitrogen) in 24- or 12-well plates, respectively.

Soluble ligands, inhibitors and other reagents. Lipopolysaccharide (LPS) was purchased from Sigma (*E. coli* 055:B5, phenol extracted). Caspase inhibitors z-YVAD, z-IEDT, Q-VD-OPH (all SM Biochemicals) were used at 50 µM, and added 30 min before stimulation of cells.

Statistical analysis. Statistical significances were tested by an ANOVA Kruskall–Wallis test and Bonferroni–Dunn post hoc correction. Significances are represented in the figures as follows: *, $P \leq 0.05$; **, $P \leq 0.01$; ***, $P \leq 0.001$. NS, not statistically significant; hash symbol, not detected.

31. Torchinsky, M. B., Garaude, J., Martin, A. P. & Blander, J. M. Innate immune recognition of infected apoptotic cells directs T_H17 cell differentiation. *Nature* **458,** 78–82 (2009).
32. Blander, J. M. & Medzhitov, R. Regulation of phagosome maturation by signals from toll-like receptors. *Science* **304,** 1014–1018 (2004).
33. Sutterwala, F. S. *et al.* Critical role for NALP3/CIAS1/Cryopyrin in innate and adaptive immunity through its regulation of caspase-1. *Immunity* **24,** 317–327 (2006).
34. Kuida, K. *et al.* Altered cytokine export and apoptosis in mice deficient in interleukin-1β converting enzyme. *Science* **267,** 2000–2003 (1995).
35. Sato, M. *et al.* Distinct and essential roles of transcription factors IRF-3 and IRF-7 in response to viruses for IFN-α/β gene induction. *Immunity* **13,** 539–548 (2000).
36. Maurelli, A. T., Baudry, B., d'Hauteville, H., Hale, T. L. & Sansonetti, P. J. Cloning of plasmid DNA sequences involved in invasion of HeLa cells by *Shigella flexneri*. *Infect. Immun.* **49,** 164–171 (1985).
37. Haraga, A., Ohlson, M. B. & Miller, S. I. Salmonellae interplay with host cells. *Nature Rev. Microbiol.* **6,** 53–66 (2008).
38. Schnupf, P. & Portnoy, D. A. Listeriolysin O: a phagosome-specific lysin. *Microbes Infect.* **9,** 1176–1187 (2007).
39. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).

40. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19,** 185–193 (2003).
41. Irizarry, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31,** e15 (2003).
42. Dai, M. *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33,** e175 (2005).
43. Saeed, A. I. *et al.* TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34,** 374–378 (2003).
44. Saeed, A. I. *et al.* TM4 microarray software suite. *Methods Enzymol.* **411,** 134–193 (2006).
45. Sartor, M. A. *et al.* Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. *BMC Bioinformatics* **7,** 538 (2006).
46. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100,** 9440–9445 (2003).
47. Samarajiwa, S. A., Forster, S., Auchettl, K. & Hertzog, P. J. INTERFEROME: the database of interferon regulated genes. *Nucleic Acids Res.* **37,** D852–D857 (2009).
48. Pahl, H. L. Activators and target genes of Rel/NF-κB transcription factors. *Oncogene* **18,** 6853–6866 (1999).
49. Coll, R. C. & O'Neill, L. A. New insights into the regulation of signalling by toll-like receptors and nod-like receptors. *J. Innate Immun.* **2,** 406–421 (2010).
50. Lim, S. Y., Bauermeister, A., Kjonaas, R. A. & Ghosh, S. K. Phytol-based novel adjuvants in vaccine formulation: 2. Assessment of efficacy in the induction of protective immune responses to lethal bacterial infections in mice. *J. Immune Based Ther. Vaccines* **4,** 5 (2006).
51. Hornung, V. *et al.* 5′-Triphosphate RNA is the ligand for RIG-I. *Science* **314,** 994–997 (2006).

LETTER

doi:10.1038/nature10073

Forces between clustered stereocilia minimize friction in the ear on a subnanometre scale

Andrei S. Kozlov[1], Johannes Baumgart[2], Thomas Risler[3,4,5], Corstiaen P. C. Versteegh[1,6] & A. J. Hudspeth[1]

The detection of sound begins when energy derived from an acoustic stimulus deflects the hair bundles on top of hair cells[1]. As hair bundles move, the viscous friction between stereocilia and the surrounding liquid poses a fundamental physical challenge to the ear's high sensitivity and sharp frequency selectivity. Part of the solution to this problem lies in the active process that uses energy for frequency-selective sound amplification[2,3]. Here we demonstrate that a complementary part of the solution involves the fluid–structure interaction between the liquid within the hair bundle and the stereocilia. Using force measurement on a dynamically scaled model, finite-element analysis, analytical estimation of hydrodynamic forces, stochastic simulation and high-resolution interferometric measurement of hair bundles, we characterize the origin and magnitude of the forces between individual stereocilia during small hair-bundle deflections. We find that the close apposition of stereocilia effectively immobilizes the liquid between them, which reduces the drag and suppresses the relative squeezing but not the sliding mode of stereociliary motion. The obliquely oriented tip links couple the mechanotransduction channels to this least dissipative coherent mode, whereas the elastic horizontal top connectors that stabilize the structure further reduce the drag. As measured from the distortion products associated with channel gating at physiological stimulation amplitudes of tens of nanometres, the balance of viscous and elastic forces in a hair bundle permits a relative mode of motion between adjacent stereocilia that encompasses only a fraction of a nanometre. A combination of high-resolution experiments and detailed numerical modelling of fluid–structure interactions reveals the physical principles behind the basic structural features of hair bundles and shows quantitatively how these organelles are adapted to the needs of sensitive mechanotransduction.

A hair bundle is a microscopic array of quasi-rigid, cylindrical stereocilia separated by small gaps filled with viscous endolymph. Like an array of organ pipes, the stereocilia vary monotonically in length across the hair bundle (Supplementary Information section 1). The tip of each short stereocilium is attached to the side of the longest adjacent stereocilium by a tip link, the tension in which controls the opening and closing of transduction channels. Adjacent stereocilia are also interconnected along all three hexagonal axes by horizontal top connectors. At the tall edge of the bundle in many species stands a single kinocilium, the process to which mechanical stimuli are applied and that is ligated to the adjacent stereocilia by kinociliary links.

When a solid object such as a hair bundle moves through a viscous fluid, the interplay between viscosity and inertia produces a spatial gradient of fluid velocity and the shear between successive layers of fluid causes friction[4]. The characteristic decay length of the shear waves created by an oscillating body scales as $\sqrt{\eta/(\omega\rho)}$, in which η is the fluid's dynamic viscosity, ρ is its density and ω is the angular frequency of motion[5]. Because this length scale greatly exceeds the distance between stereocilia, viscous forces can couple all motions within a hair bundle. On the other hand, the pivotal stiffness of individual stereociliary rootlets opposes deflection. The viscous forces in the endolymph, elastic forces in the stereociliary pivots and links, and (at high frequencies) inertial forces associated with the liquid and stereociliary masses together determine all the motions within a bundle.

Although stereociliary motion can be measured directly with an interferometer (Supplementary Information section 1), a qualitative appreciation of the liquid's movement can be obtained from the associated drag. When a fluid moves between nearby cylinders with axes perpendicular to the flow, the drag on each cylinder exceeds that on an identical cylinder placed alone in a flow with the same average velocity. At a Reynolds number well below one, this effect is strong and long-range[6,7]. One might therefore expect a drag coefficient for a hair bundle several hundred times that of an isolated stereocilium. Instead, the measured values are of similar magnitude: for six interferometric measurements in each case, the drag coefficient for a single stereocilium is 16 ± 5 nN s m^{-1}, whereas that for an entire bundle lacking tip links is only 30 ± 13 nN s m^{-1}. Because we determined the drag coefficient for hair bundles that lacked tip links and displayed coherent Brownian motion, the latter value is about a quarter of that typically reported in the literature[8]. We note that these values resemble those calculated for geometrical solids of similar dimensions pivoting at their bases and evaluated at their tips[9,10]: 14 nN s m^{-1} for a cylinder of the size of a stereocilium and 29 nN s m^{-1} for a hemi-ellipsoid with the dimensions of a hair bundle. The small difference between the drag coefficients for a single stereocilium and for an entire hair bundle reveals the striking advantage that grouping stereocilia in a tightly packed array offers to the auditory system.

Although stereocilia may slide past each other quite easily, large forces are required to squeeze them together or separate them. To estimate these forces, we constructed a macroscopic model of a hair bundle with the surrounding liquid, preserving the scaling between the physical quantities of importance (Supplementary Information section 2). A simplified model of a bullfrog's hair bundle enlarged 12,000 times was placed in a 2.2% solution of methylcellulose, which is 5,000 times as viscous as water. A single stereocilium was pulled at speeds of 0.015–1.11 mm s^{-1} while the frictional force was measured. After rescaling the time, length and mass values to those of a biological hair bundle, we estimated the drag coefficient for the small-gap separation of a single stereocilium to be 1,000–10,000 nN s m^{-1}, which is several hundred times that for the movement of an isolated stereocilium. This order-of-magnitude demonstration confirmed that very large frictional forces oppose the squeezing motion, indicating the importance of hydrodynamics in the coupling of stereocilia.

Elastic forces become dominant in the low-frequency regime and inertial forces become dominant in the high-frequency regime of hair-bundle motion. To quantify the forces as a function of frequency, we developed a finite-element model in which we could manipulate the mechanical properties of the elastic links while explicitly representing the liquid around and between the stereocilia (Supplementary

[1]Howard Hughes Medical Institute and Laboratory of Sensory Neuroscience, The Rockefeller University, 1230 York Avenue, New York, New York 10065, USA. [2]Institute of Scientific Computing, Department of Mathematics, Technische Universität Dresden, 01062 Dresden, Germany. [3]Institut Curie, Centre de Recherche, F-75005 Paris, France. [4]UPMC Université Paris 06, UMR 168, F-75005 Paris, France. [5]CNRS, UMR 168, F-75005 Paris, France. [6]Experimental Zoology Group, Wageningen University, 6709 PG Wageningen, The Netherlands.

Information section 3). The model has about 800,000 degrees of freedom and is the first finite-element model to resolve the liquid motion in the gaps between stereocilia as well as in the outer boundary layer. The hair bundle is excited in the model by imposing an oscillatory displacement at varying frequencies on the kinocilium.

First, we examined the model including only pivotal stiffness, hydrodynamic drag and inertial mass (Supplementary Movie 1). At low frequencies, the viscous force is small and only the stimulated kinocilium and its tightly joined next neighbours move (Fig. 1a). The associated drag coefficient is about 5,000 nN s m^{-1} (Fig. 1b inset), a value in agreement with the result obtained with the scaled dynamical model. Because frictional forces increase linearly with frequency whereas elastic coupling remains constant for a given displacement, hydrodynamic coupling progressively entrains the whole hair bundle at higher frequencies (Fig. 1a). As the squeezing modes subside, the drag coefficient per stereocilium decreases, dropping by two orders of magnitude by 100 Hz (Fig. 1b). Above that frequency the entire bundle moves as a unit (Fig. 1a).

Exciting the hair bundle and recording the linear responses at its opposite edges allowed us to compute the coherence of motion, a quantity that could be directly compared with interferometric measurements[11] (Supplementary Information sections 1, 3 and 4). A hair bundle in the finite-element model without any interstereociliary linkages displays a coherence exceeding 0.6 between 100 Hz and 5 kHz until inertia intervenes at higher frequencies (Fig. 1c). Adding horizontal top connectors with a stiffness of 20 mN m^{-1} to the model strongly increases the coherence, especially at low frequencies, and reduces the drag (Fig. 1b and c and Supplementary Movie 2). This value for the stiffness of top connectors was chosen such that the output coherence spectrum matched the experimental observations. It is corroborated by the distortion-product experiments discussed below and accords with published experimental and modelling studies[12–14].

Adding to the model tip links with a stiffness of 1 mN m^{-1}, rather than top connectors, introduces some elastic coupling between the stereocilia of a given column (Fig. 1c and Supplementary Movie 3), but this coupling is inefficient. For low frequencies at which hydrodynamic coupling is weak, only the excited column moves significantly. Moreover, because they are oriented obliquely, the tip links pull the stereocilia towards one another during positive deflections and allow them to separate during the complementary half-cycles. Both effects dramatically increase the drag, which originates almost entirely from the liquid within the hair bundle (Fig. 1b).

Including both horizontal top connectors and tip links in the model increases the coherence for all frequencies below 5 kHz to 0.94 (Fig. 1c and Supplementary Movie 4), a value comparable to the experimental measurement. This model displays a low drag coefficient of

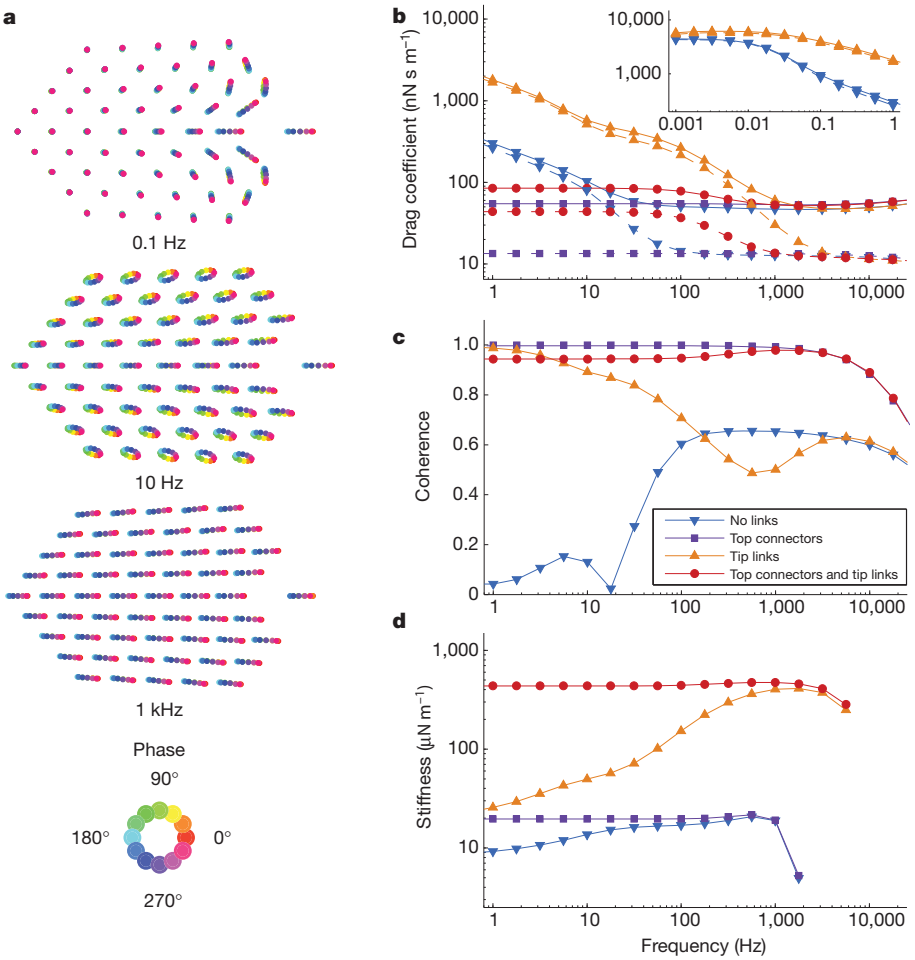


Figure 1 | Finite-element analysis of fluid–structure interactions in a hair bundle. **a**, Three top views illustrate the calculated motion of a hair bundle without elastic elements other than the kinociliary links and rootlets in response to sinusoidal deflections of the kinocilium, which lies at the right in each diagram. The colour scale (at the bottom) identifies successive positions through one cycle of stimulation with phase progressing counterclockwise. As the frequency increases, the stereocilia display a transition from weakly coupled to collective motion. The frequency dependence of the drag coefficient (**b**), the coherence (**c**) and the stiffness (**d**) are obtained from the model with four configurations of the coupling between stereocilia: with only pivotal stiffness and hydrodynamic drag (blue downtriangles); adding horizontal top connectors with a stiffness of 20 mN m^{-1} (purple squares); adding instead tip links with a stiffness of 1 mN m^{-1} (orange uptriangles); and adding both top connectors and tip links (red circles). The drag coefficient in **b** was calculated in the presence of liquid both outside and inside the hair bundle (solid lines) as well as with the liquid inside only (dashed lines). The inset in **b**, which has axis labels identical to those of the main panel, displays the behaviour of two model configurations at extremely low frequencies.

85 nN s m^{-1} that changes little with frequency (Fig. 1b), with the drag originating primarily from the external liquid but with some contribution from relative motions in the bundle, and a stiffness of 450 μN m^{-1} (Fig. 1d), similar to that reported for intact hair bundles[8,15]. We note that at frequencies below 1 kHz the tip links strongly increase the hair bundle's drag, whereas the top connectors largely suppress this effect. At higher frequencies, the liquid alone provides such a strong coupling that the tip links do not affect the drag significantly. This frequency-dependent transition between elastic and viscous regimes might explain why some high-frequency hair cells, in particular mammalian inner hair cells, apparently lack top connectors[16].

We next explored the fluid–structure interactions in an analytically tractable and intrinsically stochastic model that allowed us to generate time series that could be compared directly with experiments (Supplementary Information section 5). Unlike the harmonic single-point excitation in the finite-element model, the movement in this instance was caused by the coupling of each individual stereocilium to the thermal bath through a Langevin equation. The movements between each pair of stereocilia were derived from a basis set of elementary motions, for which we solved the Stefan–Reynolds equations within the lubrication approximation (Supplementary Information section 6).

Setting the elastic coupling to zero, we obtained a damping matrix with eigenvalues spanning about three orders of magnitude from the least damped collective modes to the most damped relative ones (Fig. 2a). This analysis shows that drag values that are low and comparable to those measured experimentally arise only when the common modes predominate. We next simulated stereociliary motions that matched the experimental records in time resolution and computed the associated coherence, which exceeded 0.95 up to 5 kHz (Fig. 2b). Changing the elastic coupling in the model revealed its importance at low frequencies, whereas viscous coupling intervened at higher frequencies.

These results show that the magnitude of the relative motion in a hair bundle depends on the balance between hydrodynamic and elastic forces. That hair bundles undergoing Brownian motion display a high coherence[11] indicates that the relative mode is very small, which makes it difficult to detect and quantify. We therefore devised an experiment in which hair bundles were stimulated at physiological amplitudes to evoke channel gating and cause intrinsic oscillations at the combination frequencies (Supplementary Information section 7). Because the gating of each mechanotransduction channel in a hair bundle changes

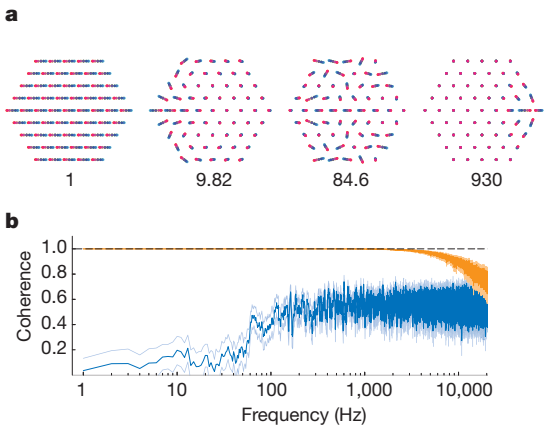


Figure 2 | Fluid–structure interactions in a stochastic model. **a**, Calculations for a model with only pivotal stiffness and hydrodynamic drag, two degrees of freedom per stereocilium, and no kinocilium yield 122 eigenmodes, of which four representative examples are shown. The eigenmodes of the damping matrix progress from a collective mode with a low-drag eigenvalue to a relative mode that is a thousand times as dissipative. The reported eigenvalues are expressed in multiples of the smallest one. **b**, The calculated coherence of motion for a hair bundle with a top-connector stiffness of 20 mN m^{-1} (orange) or 20 μN m^{-1} (blue) illustrates the importance of elastic linkages at low frequencies and of viscous coupling at high frequencies.

the force in the associated tip link[17], it must cause a relative motion of the interconnected stereocilia that is balanced by the frictional drag and elastic linkages. Blocking the distortion products at one edge of the hair bundle while measuring the relative motion at the opposite edge allowed us to isolate and quantify the amount of splay between adjacent stereocilia during small deflections, assess the forces at play and compare the results with our model.

In agreement with a previous report[18], using a flexible glass probe attached to a hair bundle's tall edge to stimulate it at two frequencies evoked distortion products at several combination frequencies. These distortion products were robust at both edges of a hair bundle and disappeared when the tip links were disrupted by 1,2-bis(*o*-aminophenoxy) ethane-*N*,*N*,*N*′,*N*′-tetraacetic acid (BAPTA), confirming that the distortion was caused by the gating of mechanotransduction channels. We then used a stiff glass probe to stimulate the long edge of the hair bundle. The rigid probe in this key experiment prevented any internally generated motion from contaminating the signal at the tall edge, which therefore consisted purely of the two excitation frequencies. With this constraint, the distortion products were significant only at the free, short edge of the hair bundle (Fig. 3a).

We related the distortion of the short-edge motion to the linear displacement by a power series. The inverse of the quadratic term of this fit was 0.14 ± 0.12 μm ($n = 8$) for the flexible probe and 1.6 ± 0.9 μm ($n = 4$) for the stiff probe. The distortions were therefore reduced to less than a tenth of their original value when the bundle's tall edge was forced to follow the stimulus signal exactly. The finite-element model with viscous coupling replicated this effect, with the top-connector stiffness determined independently from the other experimental data. The remaining distortions revealed that the relative movement between adjacent stereocilia was less than a nanometre, only a few times the size of a water molecule (Fig. 3a).

To confirm further the correspondence between experiment and modelling, we tested the prediction that removal of the horizontal top connectors should diminish the coherence (Fig. 1c) and increase the overall drag (Figs 1b and 2a). We placed hair bundles in a Ca^{2+}-free, iso-osmotic solution of mannitol having the same viscosity as

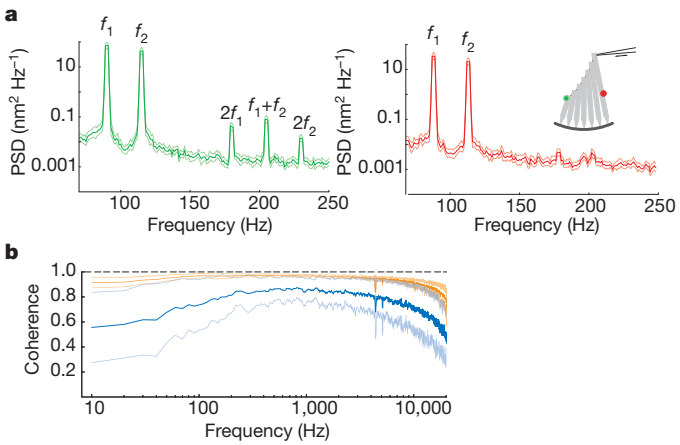


Figure 3 | Experimental verification of model predictions. **a**, Power spectra reveal that exciting a hair bundle with a stiff glass probe at two frequencies ($f_1 = 90$ Hz and $f_2 = 115$ Hz) generates distortion products marked by peaks of power-spectral density (PSD) at the second harmonics ($2f_1 = 180$ Hz and $2f_2 = 230$ Hz) and at the combination frequency ($f_1 + f_2 = 205$ Hz). Because the stiff probe suppresses internally generated movements at the tall edge (right panel), the distortion products are present only at the free short edge (left panel). The presence of distortion products directly demonstrates the relative mode of motion within the array. The schematic diagram of a hair bundle in the inset indicates the stimulating probe attached at the bundle's top and the positions of the red and green laser spots used in the interferometric measurements. **b**, The coherence in perilymph (orange) declines appreciably in the presence of mannitol (blue), which disrupts the horizontal top connectors. The mean values are accompanied by 95% confidence intervals in light orange and light blue, respectively.

saline solution but a lower ionic strength. This medium has been reported to remove the top connectors[19], and we verified the treatment's effect by transmission electron microscopy (Supplementary Information section 8). After 20 min of treatment, the top connectors were overstretched or broken, but not entirely absent (data not shown). Some elastic coupling thus persisted in mannitol. As our model predicted, the procedure decoupled the stereocilia and increased the drag, although quantitatively the effect was variable from cell to cell, presumably because of heterogeneity in the residual top connectors. For the same six cells in both conditions, the coherence between 100 Hz and 5 kHz declined from 0.96 ± 0.01 in perilymph to 0.83 ± 0.12 in mannitol (Fig. 3b). At the same time, the drag coefficient in mannitol increased to $99 \pm 63\ \mathrm{nN\,s\,m^{-1}}$. Together with the close match between the coherence values in the experiment and in the models, this and the results above confirm the accuracy of the numerical models and indicate that they capture the essential physics of the fluid–structure interactions in a hair bundle.

In conclusion, because all stereocilia and the liquid between them move in unison over the whole auditory spectrum, with the relative motions apparent only on a sub-nanometre scale, most stereocilia inside the hair bundle are shielded from the external liquid and experience little viscous drag. Although viscous forces might be thought to impair sensitivity and frequency selectivity, the hair bundle's structure actually minimizes energy dissipation, making it easier for the active process to keep the ear tuned. The tight clustering of stereocilia even transforms liquid viscosity into an asset by using it as a simple means of activating numerous mechanosensitive ion channels in concert.

METHODS SUMMARY

The methods used in this study are described in the Supplementary Information. Force measurements on a scaled hair-bundle model respected the physiological character of the liquid flow. The finite-element method provided approximate solutions to partial differential equations reflecting the hair bundle's geometry. The small amplitudes of motion allowed the elimination of nonlinear terms. The velocity variable of the liquid was replaced with the time derivative of the displacement; fluid pressure was approximated by linear shape functions and the displacements of liquid and solid were approximated by quadratic functions. The hydrodynamic forces between stereocilia were estimated analytically by solving the Stefan–Reynolds equations under the lubrication approximation, which is valid when the gaps between adjacent stereocilia are much smaller than their diameter. Stochastic simulations based on these results were performed for a system of linearly coupled dynamic variables, following a Langevin description with Gaussian white noise at room temperature. The integration procedure was validated by choosing time steps small enough to ensure that the results were independent of the increment. The robustness of our conclusions was investigated by a detailed parameter-variation study. We tested the effects of inertia and of the estimated top-connector stiffness and confirmed the validity of our conclusions for mammalian hair bundles.

Dual-beam differential interferometry was used to record stereociliary motions with sub-nanometre spatial and sub-millisecond temporal resolution. Fourier analysis of the records was performed with the multitaper method to obtain coherence spectra as well as stiffness and drag coefficients. Distortion products were evoked by stimulating hair bundles with calibrated glass probes. These results were used to verify the predictions of the numerical model and to measure the relative mode of motion between stereocilia directly. Transmission and scanning electron microscopy was performed by standard techniques with minor modifications.

Received 5 August 2010; accepted 24 March 2011.
Published online 22 May 2011.

1. Hudspeth, A. J. How the ear's works work. *Nature* **341,** 397–404 (1989).
2. Hudspeth, A. J. Making an effort to listen: mechanical amplification in the ear. *Neuron* **59,** 530–545 (2008).
3. Hudspeth, A. J., Jülicher, F. & Martin, P. A critique of the critical cochlea: Hopf—a bifurcation—is better than none. *J. Neurophysiol.* **104,** 1219–1229 (2010).
4. Stokes, G. G. On the effect of the internal friction of fluids on the motion of pendulums. *Trans. Camb. Phil. Soc.* **9,** 1–86 (1850).
5. Batchelor, G. K. *An Introduction to Fluid Dynamics* 353–364 (Cambridge University Press, 2000).
6. Tamada, K. & Fujikawa, H. The steady two-dimensional flow of viscous liquid at low Reynolds numbers passing through an infinite row of equal parallel circular cylinders. *Q. J. Mech. Appl. Math.* **10,** 425–432 (1957).
7. Yeom, J., Agonafer, D. D., Han, J.-H. & Shannon, M. A. Low Reynolds number flow across an array of cylindrical microposts in a microchannel and figure-of-merit analysis of micropost-filled microreactors. *J. Micromech. Microeng.* **19,** doi:10.1088/0960-1317/19/6/065025 (2009).
8. Denk, W., Webb, W. W. & Hudspeth, A. J. Mechanical properties of sensory hair bundles are reflected in their Brownian motion measured with a laser differential interferometer. *Proc. Natl Acad. Sci. USA* **86,** 5371–5375 (1989).
9. Perrin, F. Mouvement brownien d'un ellipsoïde (I). Dispersion diélectrique pour des molécules ellipsoïdales. *J. Phys. Radium* **5,** 497–511 (1934).
10. Broersma, S. Viscous force and torque constant for a cylinder. *J. Chem. Phys.* **74,** 6989–6990 (1981).
11. Kozlov, A. S., Risler, T. & Hudspeth, A. J. Coherent motion of stereocilia assures the concerted gating of hair-cell transduction channels. *Nature Neurosci.* **10,** 87–92 (2007).
12. Pickles, J. O. A model for the mechanics of the stereociliar bundle on acousticolateral hair cells. *Hear. Res.* **68,** 159–172 (1993).
13. Cotton, J. & Grant, W. Computational models of hair cell bundle mechanics: II. Simplified bundle models. *Hear. Res.* **197,** 105–111 (2004).
14. Karavitaki, K. D. & Corey, D. P. Sliding adhesion confers coherent motion to hair cell stereocilia and parallel gating to transduction channels. *J. Neurosci.* **30,** 9051–9063 (2010).
15. Martin, P., Mehta, A. D. & Hudspeth, A. J. Negative hair-bundle stiffness betrays a mechanism for mechanical amplification by the hair cell. *Proc. Natl Acad. Sci. USA* **97,** 12026–12031 (2000).
16. Verpy, E. *et al.* Stereocilin connects outer hair cells stereocilia to one another and to the tectorial membrane. *J. Comp. Neurol.* **519,** 194–210 (2011).
17. Howard, J. & Hudspeth, A. J. Compliance of the hair bundle associated with gating of mechanoelectrical transduction channels in the bullfrog's saccular hair cell. *Neuron* **1,** 189–199 (1988).
18. Jaramillo, F., Markin, V. S. & Hudspeth, A. J. Auditory illusions and the single hair cell. *Nature* **364,** 527–529 (1993).
19. Neugebauer, D. C. & Thurm, U. Surface charges of the membrane and cell adhesion substances determine the structural integrity of hair bundles from the inner ear of fish. *Cell Tissue Res.* **249,** 199–207 (1987).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. J. Hinterwirth for assistance in constructing the interferometer and B. Fabella for programming the experimental software; M. Fleischer for help with programming the fluid finite-element model; R. Gärtner and A. Voigt for discussions of the finite-element model and stochastic computations; M. Lenz for discussions of stochastic computations and the analytic derivation of fluid-mediated interactions; and O. Ahmad, D. Andor and M. O. Magnasco for discussions about data analysis. This research was funded by National Institutes of Health grant DC000241. Computational resources were provided by the Center for Information Services and High Performance Computing of the Technische Universität Dresden. J.B. was supported by grants Gr 1388/14 and Vo 899/6 from the Deutsche Forschungsgemeinschaft. A.S.K. was supported by the Howard Hughes Medical Institute, of which A.J.H. is an Investigator.

Author Contributions A.S.K. organized the collaboration, designed and performed the experiments, analysed data, and wrote most of the manuscript. J.B. developed the finite-element formulation and conducted the corresponding computations, implemented the stochastic modelling, derived analytic estimates of fluid-mediated interactions, wrote the corresponding Supplementary Information sections, analysed data and edited the manuscript. T.R. derived analytic estimates of fluid-mediated interactions, developed the stochastic models, implemented the data analysis, wrote the corresponding Supplementary Information sections and edited the manuscript. C.P.C.V. built the scaled model and performed the corresponding experiment. A.J.H. designed the experiments, performed the electron microscopy and edited the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to A.J.H. (hudspaj@rockefeller.edu).

LETTER

doi:10.1038/nature10084

Telomere shortening and loss of self-renewal in dyskeratosis congenita induced pluripotent stem cells

Luis F. Z. Batista[1], Matthew F. Pech[1,2], Franklin L. Zhong[1,2], Ha Nam Nguyen[3], Kathleen T. Xie[4], Arthur J. Zaug[5], Sharon M. Crary[5], Jinkuk Choi[1,2], Vittorio Sebastiano[3,6], Athena Cherry[6], Neelam Giri[7], Marius Wernig[3,6], Blanche P. Alter[7], Thomas R. Cech[5], Sharon A. Savage[7], Renee A. Reijo Pera[2,3] & Steven E. Artandi[1,2,8]

The differentiation of patient-derived induced pluripotent stem cells (iPSCs) to committed fates such as neurons, muscle and liver is a powerful approach for understanding key parameters of human development and disease[1–6]. Whether undifferentiated iPSCs themselves can be used to probe disease mechanisms is uncertain. Dyskeratosis congenita is characterized by defective maintenance of blood, pulmonary tissue and epidermal tissues and is caused by mutations in genes controlling telomere homeostasis[7,8]. Short telomeres, a hallmark of dyskeratosis congenita, impair tissue stem cell function in mouse models, indicating that a tissue stem cell defect may underlie the pathophysiology of dyskeratosis congenita[9,10]. Here we show that even in the undifferentiated state, iPSCs from dyskeratosis congenita patients harbour the precise biochemical defects characteristic of each form of the disease and that the magnitude of the telomere maintenance defect in iPSCs correlates with clinical severity. In iPSCs from patients with heterozygous mutations in *TERT*, the telomerase reverse transcriptase, a 50% reduction in telomerase levels blunts the natural telomere elongation that accompanies reprogramming. In contrast, mutation of dyskerin (*DKC1*) in X-linked dyskeratosis congenita severely impairs telomerase activity by blocking telomerase assembly and disrupts telomere elongation during reprogramming. In iPSCs from a form of dyskeratosis congenita caused by mutations in *TCAB1* (also known as *WRAP53*), telomerase catalytic activity is unperturbed, yet the ability of telomerase to lengthen telomeres is abrogated, because telomerase mislocalizes from Cajal bodies to nucleoli within the iPSCs. Extended culture of *DKC1*-mutant iPSCs leads to progressive telomere shortening and eventual loss of self-renewal, indicating that a similar process occurs in tissue stem cells in dyskeratosis congenita patients. These findings in iPSCs from dyskeratosis congenita patients reveal that undifferentiated iPSCs accurately recapitulate features of a human stem cell disease and may serve as a cell-culture-based system for the development of targeted therapeutics.

Patients with dyskeratosis congenita have high rates of bone marrow failure, pulmonary fibrosis and cancer, and a triad of epidermal findings, including oral leukoplakia, nail dystrophy and abnormal skin pigmentation[7,11]. The severity of dyskeratosis congenita and its age of onset vary widely; the reason for this range of phenotypes is unclear, but it depends in part on the mode of inheritance and the specific genes involved. Patients with X-linked dyskeratosis congenita due to mutations in *DKC1* typically present in early childhood with the classic manifestations of the disease[11,12]. In contrast, autosomal dominant dyskeratosis congenita due to mutations in *TERT* or *TERC*, the telomerase RNA component, presents in adolescence or young adulthood, and disease manifestations are often milder, with patients commonly lacking the epidermal triad. Patients with autosomal recessive dyskeratosis congenita due to *TCAB1* mutations have the classic and severe form of the disease, with early age of onset and shortened life expectancy[13]. All forms of dyskeratosis congenita are associated with very short telomeres in peripheral blood lymphocytes[14]. Telomerase is restricted in its expression in many tissues to stem cells and progenitor cells, and the challenges in isolating and studying these rare cells have precluded a direct analysis of telomere maintenance mechanisms in stem cells from patients with dyskeratosis congenita. In skin fibroblasts, telomerase expression is silenced, but during reprogramming the *TERT* gene is reactivated and telomerase activity is reconstituted[1,15–17]. Dyskeratosis congenita iPSCs have been used to study telomerase reactivation and TERC regulation during reprogramming, but thus far disease-specific iPSCs have not recapitulated telomere shortening[15].

To study dyskeratosis congenita in patient-derived iPSCs, fibroblasts from five patients carrying different mutations in *TERT* (P704S and R979W), *TCAB1* (H376Y/G435R) and *DKC1* (L54V and ΔL37) were transduced with retroviruses or lentiviruses expressing the reprogramming factors *SOX2*, *c-Myc* (also known as *MYC*), *KLF4* and *OCT4* (Supplementary Tables 1 and 2). Dyskeratosis congenita fibroblasts were resistant to reprogramming in ambient oxygen, but successful reprogramming was achieved under low oxygen conditions (5% O_2), a method that mitigates cellular stress responses[18] (Supplementary Table 1). To generate isogenic iPSCs with the DKC1(ΔL37) mutation but with long telomeres, we reprogrammed DKC1(ΔL37) fibroblasts in which TERT and TERC were stably overexpressed (TT), which bypasses the effects of the dyskerin mutation[19] (DKC1(ΔL37/TT) iPSCs). The resulting iPSCs from dyskeratosis congenita patients were morphologically indistinguishable from human embryonic stem (ES) cells, were positive for all markers of pluripotency tested and gave rise to cells derived from all three germ layers (Supplementary Figs 1–6).

Both autosomal dominant *TERT*-mutation-positive patients presented with bone marrow failure and short telomeres, but lacked the epidermal triad (Fig. 1a, Supplementary Fig. 7a and Supplementary Table 2). To assess the effects of the mutations on telomerase catalytic activity, wild-type or mutant TERT proteins were assembled into telomerase in human 293T cells. Following immunopurification, telomerase activity of each reconstituted enzyme was analysed using a quantitative direct enzymatic assay (Fig. 1b). For each mutant TERT, the enzymatic activity of reconstituted telomerase was reduced by 90%, and the defect was not suppressed by the telomere-binding proteins POT1 and TPP1, which enhance telomerase processivity[20] (Supplementary Fig. 8). In the iPSCs derived from these patients, *TERT* messenger RNA and TERC were upregulated similarly compared with wild-type iPSCs by polymerase chain reaction with reverse transcription (RT–PCR) and northern blot, respectively (Fig. 1c). Both dyskerin and TCAB1 were strongly upregulated by western blot with reprogramming

[1]Department of Medicine, Stanford University School of Medicine, Stanford, California 94305, USA. [2]Cancer Biology Program, Stanford University School of Medicine, Stanford, California 94305, USA. [3]Institute for Stem Cell Biology & Regenerative Medicine, Department of Obstetrics and Gynecology, Stanford University School of Medicine, Stanford, California 94305, USA. [4]Department of Biochemistry, Stanford University School of Medicine, Stanford, California 94305, USA. [5]Department of Chemistry and Biochemistry, Howard Hughes Medical Institute, University of Colorado, Boulder, Colorado 80309, USA. [6]Department of Pathology, Stanford University School of Medicine, Stanford, California 94305, USA. [7]Clinical Genetics Branch, Division of Cancer Epidemiology & Genetics, National Cancer Institute, National Institutes of Health, Department of Health & Human Services, Bethesda, Maryland 20852, USA. [8]The Glenn Laboratories for the Biology of Aging, Stanford University School of Medicine, Stanford, California 94305, USA.

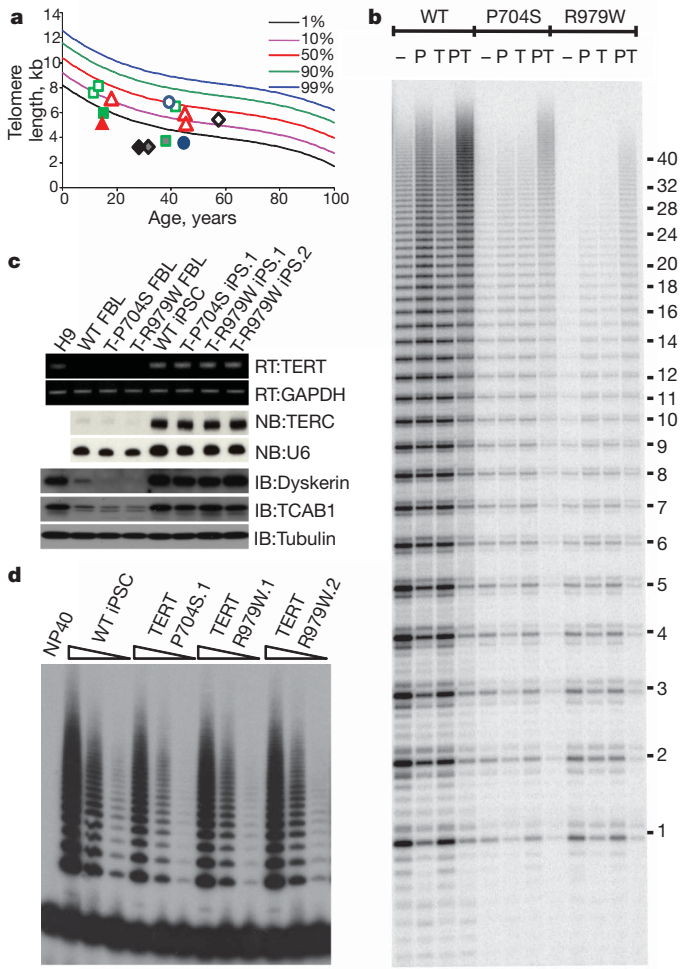


Figure 1 | Dyskeratosis congenita iPSCs with heterozygous TERT mutations show reduced telomerase levels. **a**, Telomere lengths by flow-FISH in peripheral blood lymphocytes from dyskeratosis congenita patients and their first-degree relatives. Squares, TERT(P704S) family; diamonds, TERT(R979W) family; triangles, TCAB1(H*Y/G*R) family; circles, DKC1(L54V) family. Filled symbols indicate probands; grey, carriers; open, first-degree relatives. **b**, Direct telomerase assays on wild-type (WT) TERT, or TERT mutants, assembled with TERC, with or without recombinant Pot1 (P), TPP1 (T) or PT. Numbers along the right-hand side indicate the number of telomeric repeats synthesized. **c**, Expression of TERT, TERC, DKC1 and TCAB1 with reprogramming. FBL, fibroblast; IB, immunoblot; NB, northern blot; RT, RT–PCR; T, TERT. GAPDH, U6 and Tubulin, loading controls. Different iPSC clones are shown (referred to throughout as .1 and .2). **d**, Telomerase activity by TRAP in wild-type, TERT(P704S) and TERT(R979W) iPSCs. Range of concentrations represent fourfold serial dilutions. NP40, buffer control.

(Fig. 1c). Telomerase activity in both *TERT*-mutant iPSCs was reduced by approximately 50% compared to wild-type iPSCs, consistent with our findings that each mutant TERT protein retains only 10% residual activity, which when added to the activity from the wild-type allele would be predicted to yield 55% total activity in a heterozygote (Fig. 1d). Thus, our findings in *TERT*-mutant iPSCs are compatible with a mechanism of telomerase haploinsufficiency, whereby a 50% reduction in activity is the cause of disease in this form of dyskeratosis congenita[21,22].

The patient with compound heterozygous mutations in *TCAB1* presented with classical symptoms of dyskeratosis congenita, including very short telomeres (Fig. 1a, Supplementary Fig. 7b and Supplementary Table 2). TERT, TERC and dyskerin were each appropriately upregulated in *TCAB1*-mutant iPSCs, whereas TCAB1 protein levels were markedly reduced (Fig. 2a)[13]. Although patients with mutations in *TCAB1* have short telomeres, telomerase activity was unperturbed by *TCAB1* mutations and indistinguishable from activity in wild-type iPSCs (Fig. 2b). TCAB1 is enriched in Cajal bodies, nuclear sites

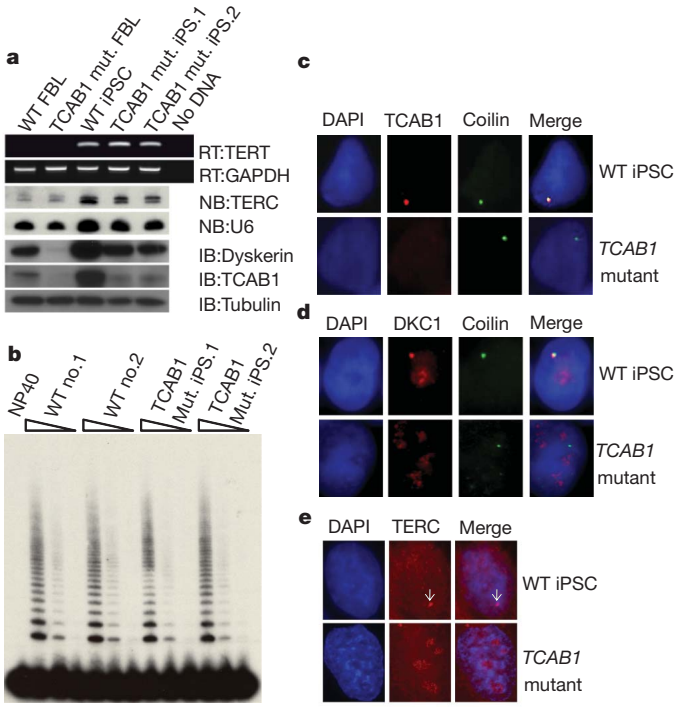


Figure 2 | Preserved activity, but pronounced mislocalization of telomerase in *TCAB1*-mutant iPSCs. **a**, Expression of TERT, TERC, DKC1 and TCAB1 with reprogramming. IB, immunoblot; NB, northern blot; RT, RT–PCR. GAPDH, U6 and Tubulin, loading controls. **b**, Telomerase activity by TRAP in wild-type and *TCAB1*-mutant iPSCs. Range of concentrations represent fourfold serial dilutions. Mut., mutation positive (H*Y/G*R) TCAB1 cells; NP40, buffer control. **c**, Immunofluorescence for TCAB1 (red) and p80-coilin (green) in wild-type and TCAB1(H*Y/G*R) iPSCs. **d**, Co-staining for dyskerin (red) and p80-coilin (green) in wild-type and TCAB1(H*Y/G*R) iPSCs. **e**, RNA FISH analysis for TERC (red) in wild-type and TCAB1(H*Y/G*R) iPSCs. White arrows, Cajal bodies. Blue, 4′,6-diamidino-2-phenylindole (DAPI).

of ribonucleoprotein modification and assembly, and is required for trafficking of telomerase to Cajal bodies[23,24]. Whereas TCAB1 colocalized in discrete foci with the Cajal-body marker p80-coilin in wild-type iPSCs, *TCAB1*-mutant iPSCs showed a marked reduction of TCAB1 accumulation in Cajal bodies (Fig. 2c and Supplementary Fig. 9a; $P < 0.0001$). Dyskerin normally accumulates both in Cajal bodies, where it binds small Cajal-body-specific RNAs (scaRNAs) and TERC, and in the nucleolus, where it binds small nucleolar RNAs (snoRNAs) that possess an H/ACA sequence motif. Efficient accumulation of dyskerin in Cajal bodies requires functional TCAB1 (ref. 13). *TCAB1*-mutant iPSCs showed a significant reduction in dyskerin accumulation in Cajal bodies, whereas nucleolar localization of dyskerin was unperturbed (Fig. 2d and Supplementary Fig. 9b; $P < 0.0001$). RNA fluorescent *in situ* hybridization (FISH) revealed that, whereas TERC localized to a single Cajal body focus in wild-type iPSCs, it showed marked mislocalization to nucleoli in *TCAB1*-mutant iPSCs (Fig. 2e and Supplementary Figs 9c and 10; $P < 0.0001$). Together, these data show that TCAB1 mutations in patient-derived iPSCs result in mislocalization of the telomerase complex without affecting telomerase activity. Our findings indicate that simple catalytic assays can falsely suggest that telomerase is active in a setting in which the telomerase enzyme is profoundly dysfunctional, results reminiscent of the first telomerase mutations in yeast[25].

Patients with X-linked dyskeratosis congenita included one with classic dyskeratosis congenita due to the DKC1(ΔL37) mutation[12,15] and another who presented with bone marrow failure, the epidermal triad and very short telomeres due to a DKC1(L54V) mutation (Fig. 1a, Supplementary Fig. 7c and Supplementary Table 2). *TERT* mRNA, dyskerin protein and TCAB1 protein were upregulated appropriately after cellular reprogramming in *DKC1*-mutant iPSCs (Fig. 3a). Dyskerin

serves a central role in assembling telomerase and other ribonucleoprotein complexes with RNAs containing H/ACA motifs[12,26]. The H/ACA motif within TERC is shared with scaRNAs and a subset of snoRNAs, involved in modification of splicing RNAs and ribosomal RNAs, respectively[26]. TERC was reduced in *DKC1*-mutant fibroblasts by northern blot, consistent with previous studies[12] (Fig. 3a). Despite an upregulation of TERC with reprogramming, TERC in *DKC1*-mutant iPSCs remained significantly suppressed compared with wild-type iPSC controls. DKC1 point mutations selectively reduced TERC levels without affecting H/ACA snoRNAs and scaRNAs, recapitulating results in lymphoblasts and fibroblasts[12] (Supplementary Fig. 11). In marked contrast to *TERT*-mutant and *TCAB1*-mutant iPSCs, all *DKC1*-mutant iPSC clones exhibited a severe reduction of telomerase activity, ranging from 5–15% of wild-type controls (Fig. 3b, c and Supplementary Fig. 12). Overexpression of TERT and TERC restored TERC levels by northern blot and rescued telomerase activity in DKC1(ΔL37/TT) fibroblasts and DKC1(ΔL37/TT) iPSCs by TRAP (Supplementary Fig. 12).

To assess the composition of the fully assembled telomerase holoenzyme in *DKC1*-mutant iPSCs, dyskerin and TCAB1 were immunoprecipitated from whole-cell extracts using antibodies directed against each protein. TERC was readily detected by northern blot in dyskerin and TCAB1 complexes from wild-type iPSCs. In contrast, the amount of TERC assembled with either dyskerin or TCAB1 was markedly reduced in DKC1(ΔL37) iPSCs (Fig. 3d and Supplementary Fig. 13). Overexpression of TERT and TERC in DKC1(ΔL37/TT) iPSCs rescued the assembly defect and led to an amount of TERC in the mature holoenzyme that exceeded wild-type levels. Overall the amount of TERC in dyskerin and TCAB1 complexes in *DKC1*-mutant, wild-type and DKC1(ΔL37/TT) iPSCs correlated directly with telomerase enzymatic activity in X-linked dyskeratosis congenita iPSCs. Thus, the reduction in both TERC and telomerase activity in *DKC1*-mutant iPSCs is consistent with a defect in a dyskerin-mediated assembly step, impairing the maturation of the active telomerase complex.

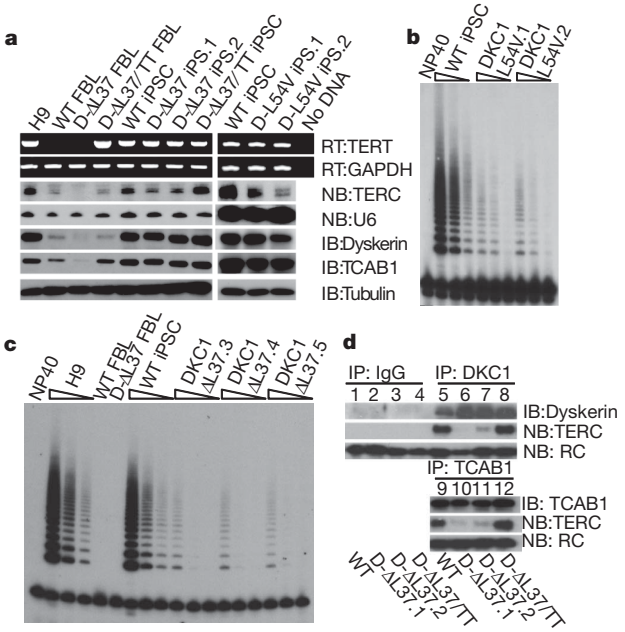


Figure 3 | Diminished TERC levels, reduced activity and impaired assembly of mature telomerase in X-linked dyskeratosis congenita iPSCs. **a**, Expression of TERT, TERC, DKC1 and TCAB1 with reprogramming. D, dyskerin. IB, immunoblot; NB, northern blot; RT, RT–PCR. GAPDH, U6 and Tubulin, loading controls. **b**, **c**, Telomerase activity by TRAP in DKC1(L54V) (**b**) and DKC1(ΔL37) (**c**) iPSCs. Range of concentrations represent fourfold serial dilutions. NP40, buffer control. Internal PCR control band at bottom of gel. **d**, Analysis of mature telomerase in iPSCs. Immunoprecipitation of 1 mg of whole-cell extracts with IgG, anti-dyskerin antibodies, or anti-TCAB1 antibodies. Purified complexes were analysed for the indicated proteins by immunoblot and for TERC by northern blot. Recovery control (RC), TERC fragment control for differential recovery of RNA.

Upregulation of telomerase leads to significant telomere lengthening during reprogramming of wild-type fibroblasts[1,15,16] (Fig. 4a–d). However, in *TERT*-mutant iPSCs, telomere elongation during reprogramming was blunted, with telomeres in *TERT*-mutant iPSCs always remaining significantly shorter than in wild-type iPSCs (Fig. 4a and Supplementary Fig. 14). In marked contrast, telomere elongation failed in all *TCAB1*-mutant iPSCs and *DKC1*-mutant iPSCS. For both *TCAB1*-mutant iPSCs and *DKC1*-mutant iPSCs, telomeres were shorter than in their parental fibroblasts and telomeres continued to shorten as cells divided in culture (Fig. 4b–d and Supplementary Figs 14 and 15). In DKC1(ΔL37/TT) fibroblasts and iPSCs, telomerase overexpression fully restored telomere elongation, with telomere lengths increasing significantly beyond those of their wild-type counterparts (Fig. 4d). These data show that telomerase mutations can severely impair telomere maintenance in dyskeratosis congenita iPSCs, providing evidence for a defect in maintaining telomeres in dyskeratosis congenita stem cells.

With extended proliferation in cell culture of DKC1(ΔL37) iPSCs, telomeres continued to shorten through passage 19 and the bulk population of telomeres reached a plateau at passage 26 by Southern blot (Supplementary Fig. 15b). Using telomere FISH, we found that telomere signals were readily detected at all chromosome ends in wild-type iPSCs and in DKC1(ΔL37/TT) iPSCs. In contrast, average telomere intensity was greatly reduced in DKC1(ΔL37) iPSCs, which also showed an increase in the number of signal-free ends (SFEs), chromosome ends lacking detectable telomere repeats (Fig. 4e–h; $P < 0.01$). Continued passage of DKC1(ΔL37) iPSCs resulted in an abrupt increase in spontaneous differentiation within iPSC colonies and the culture could no longer be maintained as undifferentiated iPSCs after passage 36. Critical telomere shortening leads to a loss of telomere capping function, which triggers a DNA damage response that activates the p53 tumour suppressor protein. The p53 pathway was strongly activated in DKC1(ΔL37) iPSCs at passage 36, as evidenced by p53 protein stabilization and induction of its downstream target p21 by western blot. No such activation of p53 was seen at passage 9, or in late-passage human ES cells (Fig. 4i). Taken together, these data show that impaired telomere maintenance in dyskeratosis congenita iPSCs can ultimately compromise self-renewal, resulting in a finite cellular lifespan.

Our data in patient-derived iPSCs provide evidence for severe defects in telomerase function and telomere maintenance in stem cells from dyskeratosis congenita patients. The spontaneous differentiation in *DKC1*-mutant iPSCs indicates that exhaustion of self-renewal in haematopoietic stem cells and other tissue stem cells may underlie the tissue defect in dyskeratosis congenita. Restoration of telomerase function through pharmacological or genetic means in stem cells from blood, lung or epidermal tissues may therefore provide a rational guide for therapy of dyskeratosis congenita. Data from these iPSCs provide an explanation for the long-standing clinical observation that X-linked dyskeratosis congenita is often more severe and presents at a younger age than autosomal dominant dyskeratosis congenita caused by mutations in *TERT* or *TERC*[8]. Our data indicate that effective telomerase activities in the 15–50% range may represent a critical threshold in which telomere maintenance is particularly impaired. Thus, a reduction in telomerase activity to the 15–50% range may be necessary to yield severe phenotypes in a single generation, whereas genetic anticipation[21,22] through inheritance of heterozygous mutations in *TERT* for several generations may be important in eliciting disease phenotypes for autosomal dominant patients with greater than 50% residual telomerase activity. Together, our data show that many important features of a human stem cell disease are accurately recapitulated in patient-derived iPSCs, providing an iPSC-based system that is not dependent upon differentiation to probe disease mechanisms or to identify potential therapeutics.

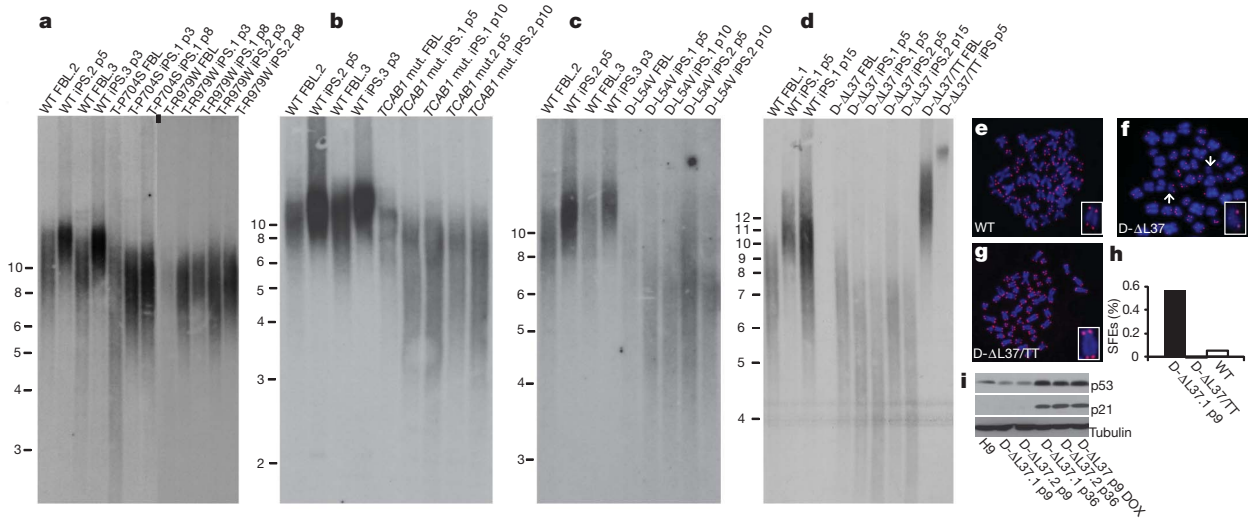


Figure 4 | Impaired telomere maintenance and loss of self-renewal in dyskeratosis congenita iPSCs. a–d, Telomere lengths by Southern blot using genomic DNA from fibroblasts and iPSCs. TERT(P704S) and TERT(R979W) iPSCs (**a**), TCAB1(H*Y/G*R) (**b**), DKC1(L54V) (**c**) and DKC1(ΔL37) and DKC1(ΔL37/TT) iPSCs (**d**) at indicated passages (p) after reprogramming. Numbers along the left-hand sides of **a**–**d** show molecular weight in kilobases. Filled black square in **a**, membrane cut for hybridization. **e**–**h**, Telomere FISH on metaphase chromosomes from wild-type iPSCs (**e**), DKC1(ΔL37) iPSC clone 1 (**f**) and DKC1(ΔL37/TT) (**g**) iPSCs at passage 22. White arrows, SFEs. High magnification, inset. **h**, Quantification of SFEs per metaphase. **i**, Western blot for p53 and p21 at passage 9 and passage 36 in DKC1(ΔL37) iPSCs. DOX, doxorubicin treated. Tubulin, loading control.

METHODS SUMMARY

TERT (P704S; R979W), *TCAB1* (H376Y/G435R) and *DKC1* (L54V) fibroblasts were obtained from skin biopsies from patients in the National Cancer Institute's Institutional Review Board-approved study, "Etiologic Investigation of Cancer Susceptibility in Inherited Bone Marrow Failure Syndromes (IBMFS)" (http://marrowfailure.cancer.gov)[8]. Detailed medical record review, physical examination, comprehensive laboratory evaluation, measurement of telomere length, genetic counselling and mutation analyses were conducted. DKC1(ΔL37) fibroblasts were purchased from Coriell Cell Repositories.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 26 February 2010; accepted 30 March 2011.
Published online 22 May 2011.

1. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131,** 861–872 (2007).
2. Rashid, S. T. *et al.* Modeling inherited metabolic disorders of the liver using human induced pluripotent stem cells. *J. Clin. Invest.* **120,** 3127–3136 (2010).
3. Liu, G. H. *et al.* Recapitulation of premature ageing with iPSCs from Hutchinson–Gilford progeria syndrome. *Nature* advance online publication doi:10.1038/nature09879 (23 February 2011).
4. Soldner, F. *et al.* Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors. *Cell* **136,** 964–977 (2009).
5. Dimos, J. T. *et al.* Induced pluripotent stem cells generated from patients with ALS can be differentiated into motor neurons. *Science* **321,** 1218–1221 (2008).
6. Maehr, R. *et al.* Generation of pluripotent stem cells from patients with type 1 diabetes. *Proc. Natl Acad. Sci. USA* **106,** 15768–15773 (2009).
7. Walne, A. J. & Dokal, I. Advances in the understanding of dyskeratosis congenita. *Br. J. Haematol.* **145,** 164–172 (2009).
8. Alter, B. P. *et al.* Malignancies and survival patterns in the National Cancer Institute inherited bone marrow failure syndromes cohort study. *Br. J. Haematol.* **150,** 179–188 (2010).
9. Lee, H. W. *et al.* Essential role of mouse telomerase in highly proliferative organs. *Nature* **392,** 569–574 (1998).
10. Allsopp, R. C., Morin, G. B., DePinho, R., Harley, C. B. & Weissman, I. L. Telomerase is required to slow telomere shortening and extend replicative lifespan of HSCs during serial transplantation. *Blood* **102,** 517–520 (2003).
11. Bessler, M., Wilson, D. B. & Mason, P. J. Dyskeratosis congenita. *FEBS Lett.* **584,** 3831–3838 (2010).
12. Mitchell, J. R., Wood, E. & Collins, K. A telomerase component is defective in the human disease dyskeratosis congenita. *Nature* **402,** 551–555 (1999).
13. Zhong, F. *et al.* Disruption of telomerase trafficking by TCAB1 mutation causes dyskeratosis congenita. *Genes Dev.* **25,** 11–16 (2011).
14. Alter, B. P. *et al.* Very short telomere length by flow fluorescence *in situ* hybridization identifies patients with dyskeratosis congenita. *Blood* **110,** 1439–1447 (2007).
15. Agarwal, S. *et al.* Telomere elongation in induced pluripotent stem cells from dyskeratosis congenita patients. *Nature* **464,** 292–296 (2010).
16. Marion, R. M. *et al.* Telomeres acquire embryonic stem cell characteristics in induced pluripotent stem cells. *Cell Stem Cell* **4,** 141–154 (2009).
17. Stadtfeld, M., Maherali, N., Breault, D. T. & Hochedlinger, K. Defining molecular cornerstones during fibroblast to iPS cell reprogramming in mouse. *Cell Stem Cell* **2,** 230–240 (2008).
18. Yoshida, Y., Takahashi, K., Okita, K., Ichisaka, T. & Yamanaka, S. Hypoxia enhances the generation of induced pluripotent stem cells. *Cell Stem Cell* **5,** 237–241 (2009).
19. Wong, J. M. & Collins, K. Telomerase RNA level limits telomere maintenance in X-linked dyskeratosis congenita. *Genes Dev.* **20,** 2848–2858 (2006).
20. Zaug, A. J., Podell, E. R., Nandakumar, J. & Cech, T. R. Functional interaction between telomere protein TPP1 and telomerase. *Genes Dev.* **24,** 613–622 (2010).
21. Vulliamy, T. *et al.* Disease anticipation is associated with progressive telomere shortening in families with dyskeratosis congenita due to mutations in *TERC*. *Nature Genet.* **36,** 447–449 (2004).
22. Armanios, M. *et al.* Haploinsufficiency of telomerase reverse transcriptase leads to anticipation in autosomal dominant dyskeratosis congenita. *Proc. Natl Acad. Sci. USA* **102,** 15960–15964 (2005).
23. Venteicher, A. S. *et al.* A human telomerase holoenzyme protein required for Cajal body localization and telomere synthesis. *Science* **323,** 644–648 (2009).
24. Tycowski, K. T., Shu, M. D., Kukoyi, A. & Steitz, J. A. A conserved WD40 protein binds the Cajal body localization signal of scaRNP particles. *Mol. Cell* **34,** 47–57 (2009).
25. Lundblad, V. & Szostak, J. W. A mutant with a defect in telomere elongation leads to senescence in yeast. *Cell* **57,** 633–643 (1989).
26. Matera, A. G., Terns, R. M. & Terns, M. P. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nature Rev. Mol. Cell Biol.* **8,** 209–220 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements L.F.Z.B. is the recipient of a Pew Fellowship. M.F.P. and K.T.X. are the recipients of NSF Graduate Research Fellowships. F.L.Z. was supported by A*STAR, Singapore. We thank the patients for their valuable contributions and L. Leathwood for study support. This work was supported, in part, by the intramural research program of the Division of Cancer Epidemiology and Genetics, NCI, NIH; by a CIRM Shared Research Laboratory grant to R.R.P.; and by grants from the NCI, NIA, NHLBI and CIRM to S.E.A.

Author Contributions L.F.Z.B., M.F.P., F.L.Z and S.E.A. designed the experiments and analysed the data; L.F.Z.B., M.F.P., F.L.Z., H.N.N., K.T.X., A.J.Z., S.M.C., J.C., V.S. and A.C. performed the experiments; N.G., M.W., B.P.A., T.R.C., S.A.S. and R.A.R.P. analysed the data; N.G., B.P.A and S.A.S. collected patients' material; L.F.Z.B. and S.E.A. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.E.A. (sartandi@stanford.edu).

METHODS

Cell culture. Human fibroblasts were cultured in fibroblast media (DMEM supplemented with 15% FBS) at 37 °C, 5% CO_2 and 5% O_2. H9 human ES cells and iPSCs were cultured on γ-irradiated mouse embryonic fibroblasts in human ES cell media consisting of DMEM F12-Glutamax, supplemented with 20% knockout serum, 0.1 mM non-essential amino-acids, 0.1 mM β-mercaptoethanol and 10 ng ml^{-1} recombinant human basic fibroblast growth factor (Invitrogen). Human ES and iPSCs were transferred to matrigel-coated plates (Invitrogen) and kept in mTeSR1 media (STEMCELL Technologies) before most experiments. In all culture conditions, human ES and iPSCs culture media was changed daily and cells were treated with collagenase (1 mg ml^{-1}, 5 min) or manually passaged every 5–6 days.

Retroviral and lentiviral production. pMXs retroviral plasmids encoding human SOX2, OCT3/4, c-Myc and KLF4 as well as the packaging vectors pUMVC and pVSV-G were purchased from Addgene. Plasmids for the production of the single polycistronic lentiviral vector were donated by G. Mostoslavsky. Retrovirus and lentivirus production for iPS cell reprogramming was performed as described[27,28]. For generation of DKC1 fibroblasts overexpressing TERT and TERC, retroviruses were generated by co-transfecting plasmids encoding RSV(Gag+Pol), VSV-G and retroviral expression plasmids containing human TERT and TERC into 293T cells using calcium phosphate precipitation. Approximately 10^5 cells were co-transduced with both genes and kept for 7 days under blasticidin and neomycin selection.

iPSC generation. For retroviral and lentiviral transduction, 10^5 dyskeratosis congenita human fibroblasts were seeded per well of a 6-well plate 24 h pre-transduction. For retroviral transduction, cells were infected for 24 h with retroviruses encoding the four reprogramming factors (5× concentration for SOX2, KLF4 and c-Myc and 10× concentration for OCT3/4 in the presence of 8 ng ml^{-1} polybrene). Cells were then washed in PBS and kept in fibroblast growth media. *DKC1* mutant fibroblasts required three rounds of viral infection on alternate days for efficient reprogramming. For lentiviral transduction, cells were kept for 24 h with single polycistronic lentiviral vectors encoding SOX2, KLF4, c-Myc and OCT3/4 (10× concentrarion, 8 ng ml^{-1} polybrene, one single round of infection), according to the protocol described previously[28]. Three days after the final round of infection, approximately 2×10^5 cells were briefly trypsinized and transferred to 10-cm dishes pre-plated with feeders. Twenty-four hours after passaging, cells were washed with PBS and human ES media was added to the plate. Media was changed every other day, until background colonies emerged, after which media was changed daily. Human ES-like colonies appeared from large background colonies 20 days after viral transduction and were manually picked on days 24–30. Colonies that maintained their ES-like morphology were further passaged and analysed for pluripotency potential. During the entire reprogramming period, cells were kept under low oxygen conditions (5% O_2). Wild-type iPSCs were reprogrammed using both retroviral and lentiviral vectors. In all figures, wild-type cells shown were reprogrammed using the same strategy as the dyskeratosis congenita cells to which they are compared. huF-4 (wild-type no. 1), huF-5 (wild-type no. 2) and huF-Q (wild-type no. 3) human dermal fibroblasts were reprogrammed and used as wild-type controls, either with retroviral (wild-type no. 1) or lentiviral vectors (wild-type no. 2 and no. 3).

Immunofluorescence. iPSCs were grown on feeders in 48-well plates, fixed with 4% paraformaldehyde/PBS, washed three times with PBS and blocked with 4% goat serum for 1 h. For OCT4 and NANOG, cells were permeabilized with 1% Triton-X/PBS for 1 h at room temperature (20–25 °C) before blocking. After blocking, primary antibodies were diluted in PBS and cells were incubated overnight at 4 °C. The following antibodies were used: SSEA3 (1:200), SSEA4 (1:200), TRA-1-60 (1:200), TRA-1-81 (1:200), all from Millipore; Nanog (1:100, Abcam), and Oct4 (1:200, Santa Cruz). After incubation, cells were washed twice with PBS and incubated with secondary antibodies for 1 h at room temperature in the dark. Secondary antibodies used were the Alexa Fluor Series from Invitrogen (all at 1:1,000). Cells where then washed three times with PBS and stained with DAPI for nuclei labelling. Images were taken using a Leica DM5000B microscope coupled to a Leica DFC360FX camera.

Telomere length analysis. Genomic DNA was collected from human fibroblasts, H9 human ES cells and iPSCs at different passages. The isolated genomic DNA was then digested with RsaI and HinfI and fractionated as described previously[29]. Membranes were prepared by Southern transfer and hybridized to a radioactively end-labelled $(TTAGGG)_4$ oligonucleotide probe as described previously[30].

Detection of telomerase activity by TRAP. Human fibroblasts and fully undifferentiated H9 ES cells and iPSCs (grown in matrigel) were lysed in NP40 buffer (25 mM HEPES-KOH, 150 mM KCl, 1.5 mM MgCl2, 10% glycerol, 0.5% NP40, and 5 mM 2-mercaptoethanol (pH 7.5) supplemented with protease inhibitors) for 15–30 min on ice. Extracts clarified by centrifugation at 16,000*g* for 10 min were quantified by Bradford assay. Telomere extension reactions were performed using 2.0 μg, 0.5 μg and 0.125 μg protein extract, and 5% of the purified telomere extension products were amplified by PCR, based on a modified protocol from the manufacturer (TRAPeze, Chemicon).

Direct telomerase activity assays. Activity of each human telomerase complex expressed in HEK 293T cells was determined by a direct assay modified from a published protocol[31]. Each human TERT contained an amino-terminal 3×Flag tag, and telomerase was immunopurified from cell extracts using anti-Flag M2 affinity gel (Sigma). The reaction mixture (20 μl) contained 1× human telomerase assay buffer (50 mM Tris-HCl, pH 8.0, 50 mM KCl, 1 mM MgCl2, 5 mM 2-mercaptoethanol, 1 mM spermidine), 0.1 μM telomeric DNA primer, 0.5 mM dATP, 0.5 mM dTTP, 2.92 μM dGTP and 0.33 μM ^{32}P-dGTP (3,000 Ci $mmol^{-1}$; 1 Ci = 37 GBq) with 6 μl of immunopurified telomerase complex on beads. Reactions were performed with telomerase alone or supplemented with human POT1, TPP1-N (amino acids 89–334), or both (0.5 μM each)[32,33]. Reactions were incubated at 30 °C for 1 h, then stopped with the addition of 100 μl of 3.6 M NH_4OAc, 20 μg of glycogen, and ethanol (500 ml). After incubating at −80 °C for 1 h, samples were centrifuged at 4 °C for 30 min. Pellets were washed with 70% ethanol and resuspended in 10 μl of H_2O followed by 10 μl of 2× loading buffer (94% formamide, 0.1× TBE, 0.1% bromophenol blue, 0.1% xylene cyanol). The heat-denatured samples were loaded onto a 10% polyacrylamide/7 M urea/1× TBE denaturing gel for electrophoresis. After electrophoresis, the gel was dried and total radioactivity incorporated into telomerase products was quantified using a Phosphorimager (GE Healthcare). A portion of each immunopurified telomerase was subjected to electrophoresis on an SDS–polyacrylamide gel and western blotting with an anti-Flag antibody to confirm equal amounts of human TERT protein in each reaction.

Telomere FISH. Metaphase chromosomes were prepared from human iPSCs by treatment with Colcemid 0.03 μg ml^{-1} KaryoMAX Colcemid Solution (Invitrogen), followed by hypotonic KCl and fixation in cold methanol-acetic acid. Chromosome hybridization with a Cy3-conjugated $(CCCTAA)_3$ peptide-nucleic acid probe has been described previously[30]. At least 10 metaphases per sample were analysed to determine the amount of signal-free ends in each sample. Differences between samples were compared by the two-tailed Fisher's exact test.

Immunoprecipitations, western blots and northern blots. Protocols for immunoprecipitation, western blotting and northern blotting have been previously described[23,34]. For western blotting, primary antibodies included Dyskerin (1:30,000 serum[35]), TCAB1 (50 ng ml^{-1}), p21 (1:400; Santa Cruz), phospho-p53-Ser15 (1:1,000; Cell Signaling) and Tubulin (1:50,000; Sigma). Generation of polyclonal antibodies has been previously described[23,34]. For northern blot analysis, total RNA was purified with Trizol reagent (Invitrogen) and treated with Turbo DNA-free kit (Ambion) to remove genomic DNA contamination. The northern blot probes and recovery control used have been described previously[23].

***In vitro* and *in vivo* differentiation.** For spontaneous *in vitro* differentiation towards endoderm, mesoderm and ectoderm fates, undifferentiated iPSCs were transferred to matrigel plates. Twenty-four hours after attaching, cells were washed with PBS and cultured thereafter with differentiation media (human ES media depleted of bFGF), changed daily. When cells reached 90% confluency, they were transferred by trypsinization to gelatin-coated 6-well plates. Twenty-one days after bFGF depletion cells were fixed with 4% paraformaldehyde/PBS, permeabilized with 1% Triton-X and incubated overnight at 4 °C with primary antibodies. The antibodies used were neuronal class III β-tubulin (TUJ1, 1:200, Covance), pan-Cytokeratin (Lu-5, 1:200, Abcam) for ectoderm staining I; α-smooth muscle actin (1:200, Abcam), Desmin (1:200, Lab Vision) for mesoderm staining; and human α-fetoprotein (1:200, R&D Biosystems) for endoderm staining. Immunofluorescence was performed as described earlier. *In vivo* differentiation through teratoma formation was done as described[1]. Briefly, approximately 10^6 iPSCs were injected subcutaneously into dorsal flanks of immunocompromised mice (a/a $Foxn1^{nu}/Foxn1^{nu}$; Jackson Laboratory). Tumours were collected after 8–10 weeks, fixed and embedded into paraffin blocks.

Genomic DNA sequencing. For genetic identification of our *TERT*, *TCAB1* and *DKC1* cells, total genomic DNA of huF4, huF5, TERT(P704S), TERT(R979W), TCAB1(H*Y/G*R), DKC1(L54V), DKC1(ΔL37) and DKC1(DL37/TT) fibroblasts, as well as their respective iPSCs was extracted by using lysis buffer[34]. One-hundred nanograms of isolated genomic DNA was used for PCR with primers flanking the specific mutations. *TERT* exon 5: Fwd: 5′-GTGGCATGAGGA TCCCGTGTGC-3′; Res: 5′-CACAGTCGGCCCCATGTGCTG-3′. *TERT* exon 12: Fwd: 5′-GGCCGTGCGAGGTTTGGATACAC-3′; Res: 5′-GTGTATCCAAA CCTCGCACGGCC-3′. *TCAB1* exon 7: Fwd: 5′- GGTCCTTTGGGAGGATAG ATGTGG-3′; Res: 5′-GGAACAGGACCTGGAGTCACCC-3′. *TCAB1* exon 9: Fwd: 5′-GTGCTGGGATCTCCGGCAGTC-3′; Res: 5′-CTGACCGGAGGCA GTGGCC-3′. *DKC1* exon 3: Fwd: 5′-GTTCAAAATCGGGTGGGAAG-3′; Res: 5′-CCAAAGTCAAGGATGCCAG-3′. After gel extraction, PCR products were sequenced. Resulting sequences were aligned by Clustal-W2 (EMBL-EBI).

G-band analysis. After mitotic arrest, monolayer cell cultures in log-phase growth were harvested by standard cytogenetic methods of trypsin dispersal, hypotonic shock and fixation. Mitotic cell slide preparations were analysed by the GTW (G-banding with trypsin and Wright's stain) banding method[36].

Bisulphite sequencing. Genomic DNA (1 μg) was treated with MethylEasy Xceed (Human Genetic Signatures) according to the manufacturer's recommendations.

The human OCT3/4 promoter was PCR-amplified and TOPO-cloned, with at least ten clones from each sample sequenced, following the protocol described[37]. **PCR.** Reverse transcription was carried with Superscript II (Invitrogen). PCR was performed using Taq DNA-polymerase (Qiagen) in a GeneAmp PCR System 9700 machine (Applied Biosystems). Primers for pluripotency (TERT, endo-OCT3/4, endo-SOX2, endo-c-Myc, NANOG, REX1, FGF4, Nodal, GDF3 and ESG1) were described[1].

27. Byrne, J. A., Nguyen, H. N. & Reijo Pera, R. A. Enhanced generation of induced pluripotent stem cells from a subpopulation of human fibroblasts. *PLoS ONE* **4,** e7118 (2009).
28. Sommer, C. A. *et al.* Induced pluripotent stem cell generation using a single lentiviral stem cell cassette. *Stem Cells* **27,** 543–549 (2009).
29. Tomlinson, R. L. *et al.* Telomerase reverse transcriptase is required for the localization of telomerase RNA to Cajal bodies and telomeres in human cancer cells. *Mol. Biol. Cell* **19,** 3793–3800 (2008).
30. Middleman, E. J., Choi, J., Venteicher, A. S., Cheung, P. & Artandi, S. E. Regulation of cellular immortalization and steady-state levels of the telomerase reverse transcriptase through its carboxy-terminal domain. *Mol. Cell. Biol.* **26,** 2146–2159 (2006).
31. Cristofari, G. & Lingner, J. Telomere length homeostasis requires that telomerase levels are limiting. *EMBO J.* **25,** 565–574 (2006).
32. Lei, M., Zaug, A. J., Podell, E. R. & Cech, T. R. Switching human telomerase on and off with hPOT1 protein in vitro. *J. Biol. Chem.* **280,** 20449–20456 (2005).
33. Wang, F. *et al.* The POT1–TPP1 telomere complex is a telomerase processivity factor. *Nature* **445,** 506–510 (2007).
34. Venteicher, A. S., Meng, Z., Mason, P. J., Veenstra, T. D. & Artandi, S. E. Identification of ATPases pontin and reptin as telomerase components essential for holoenzyme assembly. *Cell* **132,** 945–957 (2008).
35. Mochizuki, Y., He, J., Kulkarni, S., Bessler, M. & Mason, P. J. Mouse dyskerin mutations affect accumulation of telomerase RNA and small nucleolar RNA, telomerase activity, and ribosomal RNA processing. *Proc. Natl Acad. Sci. USA* **101,** 10756–10761 (2004).
36. Seabright, M. Rapid banding technique for human chromosomes. *Lancet* **2,** 971–972 (1971).
37. Deb-Rinker, P., Ly, D., Jezierski, A., Sikorska, M. & Walker, P. R. Sequential DNA methylation of the Nanog and Oct-4 upstream regions in human NT2 cells during neuronal differentiation. *J. Biol. Chem.* **280,** 6257–6260 (2005).

LETTER

doi:10.1038/nature10096

A hydrothermal origin for isotopically anomalous cap dolostone cements from south China

Thomas F. Bristow[1]†, Magali Bonifacie[1,2], Arkadiusz Derkowski[3], John M. Eiler[1] & John P. Grotzinger[1]

The release of methane into the atmosphere through destabilization of clathrates is a positive feedback mechanism capable of amplifying global warming trends that may have operated several times in the geological past[1–3]. Such methane release is a hypothesized cause or amplifier for one of the most drastic global warming events in Earth history, the end of the Marinoan 'snowball Earth' ice age, ~635 Myr ago[4–7]. A key piece of evidence supporting this hypothesis is the occurrence of exceptionally depleted carbon isotope signatures ($\delta^{13}C_{PDB}$ down to −48‰; ref. 8) in post-glacial cap dolostones (that is, dolostone overlying glacial deposits) from south China; these signatures have been interpreted as products of methane oxidation at the time of deposition[5,6,8]. Here we show, on the basis of carbonate clumped isotope thermometry, $^{87}Sr/^{86}Sr$ isotope ratios, trace element content and clay mineral evidence, that carbonates bearing the ^{13}C-depleted signatures crystallized more than 1.6 Myr after deposition of the cap dolostone. Our results indicate that highly ^{13}C-depleted carbonate cements grew from hydrothermal fluids and suggest that their carbon isotope signatures are a consequence of thermogenic methane oxidation at depth. This finding not only negates carbon isotope evidence for methane release during Marinoan deglaciation in south China, but also eliminates the only known occurrence of a Precambrian sedimentary carbonate with highly ^{13}C-depleted signatures related to methane oxidation in a seep environment. We propose that the capacity to form highly ^{13}C-depleted seep carbonates, through biogenic anaerobic oxidation of methane using sulphate, was limited in the Precambrian period by low sulphate concentrations in sea water[9]. As a consequence, although clathrate destabilization may or may not have had a role in the exit from the 'snowball' state, it would not have left extreme carbon isotope signals in cap dolostones.

A common sedimentary motif marking the end of the severe Marinoan ice age (~635 Myr ago[10]) is observed in rocks on almost all of the present day continents. Glaciogenic deposits, formed at equatorial latitudes in some places, are sharply overlain by metre-scale intervals of dolostone[11]. These cap dolostones contain enigmatic sedimentary structures and unusual stable carbon and sulphur isotope signatures hypothesized to reflect climate change associated with deglaciation[11]. Proposed drivers of global warming during this period include deep-ocean CO_2 outgassing during post-glacial ocean overturn[12], ice–albedo feedback[11] or methane release caused by the destabilization of clathrates[4].

The lithologies and sedimentary structures observed in the cap dolostone of south China, which forms the basal 3–5 m of the Doushantuo Formation, are typical of cap dolostones worldwide[6,11] (Supplementary Discussion). But unlike other cap dolostones that have mildly negative $\delta^{13}C_{PDB}$ values (−2 to −4‰; ref. 11), three sections in the Yangtze Gorges area host highly ^{13}C-depleted calcites ($\delta^{13}C_{PDB}$ down to −48‰; ref. 8). Petrographic textures of the highly ^{13}C-depleted calcites

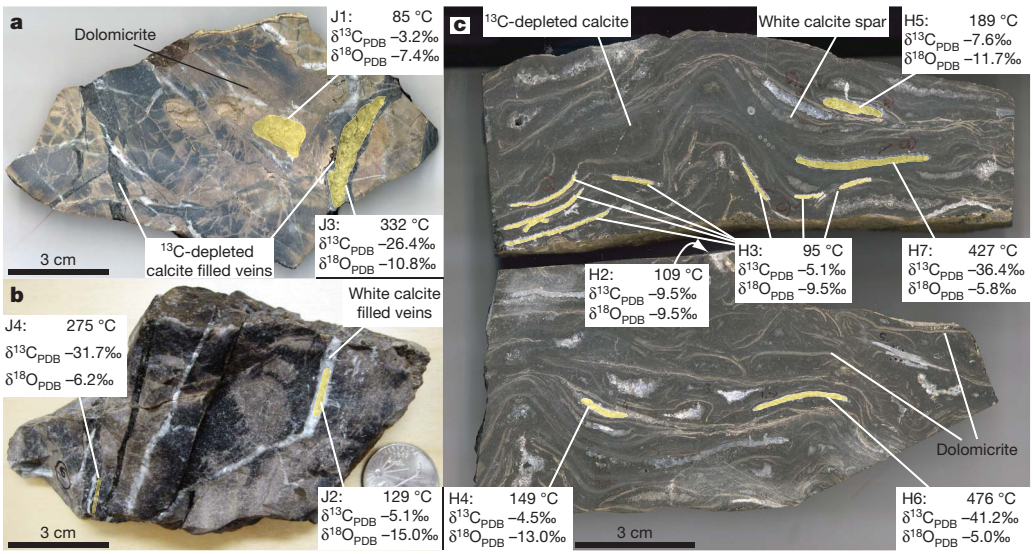


Figure 1 | Crystallization temperatures (based on Δ_{47} measurements) and C and O stable isotope data from various carbonate phases in two samples of the Doushantuo Formation cap dolostone. a, **b**, Two views of sample 1, collected from the lower unit of the cap at Jiulongwan (Fig. 3; Supplementary Fig. 1). **c**, Two pieces of sample 3, collected from the middle of the cap dolostone at Huajipo (Fig. 3; Supplementary Fig. 1). The highest temperatures come exclusively from calcite cements with exceptionally low $\delta^{13}C_{PDB}$ values. Areas drilled for isotopic analysis are highlighted in yellow (Supplementary Fig. 2 shows unmarked images of the samples) and labelled (J*x* and H*y* refer to sample spots from Jiulongwan and Huajipo, respectively) for cross-referencing with data in Supplementary Table 1. The temperatures and isotope values shown are the mean of two or three replicate measurements of the same powder.

[1]Division of Geological and Planetary Sciences, California Institute of Technology, 1200 East California Boulevard, Pasadena, California 91125, USA. [2]Équipe de Géochimie des Isotopes Stables, Institut de Physique du Globe de Paris, Sorbonne Paris Cité, Université Paris Diderot, UMR 7154 CNRS, F-75005 Paris, France. [3]Institute of Geological Sciences, Polish Academy of Sciences, Senacka 1, 31-002 Kraków, Poland. †Present address: NASA Ames Research Center, Moffett Field, California 94035, USA.

and associated carbonates have been likened to modern methane seeps[5,6,8]. Additionally, $\delta^{18}O_{PDB}$ values as high as −4‰ in some of the most ^{13}C-depleted calcites are close to values expected in equilibrium with sea water at Earth-surface temperatures, and have been used to support the hypothesis that methane oxidation occurred at the time of deposition[5].

To test this interpretation, we examined the paragenetic history of three representative samples that contain, amongst other phases, highly ^{13}C-depleted calcite. The samples were collected from the lower and middle units of the cap dolostone from two sections of the Doushantuo Formation in the Yangtze Gorges area at Jiulongwan and Huajipo (see below; Supplementary Fig. 1). Samples consist of: (1) early dolomicrite; (2) highly ^{13}C-depleted grey calcite, with crystals up to 1 mm in length, filling fractures that cross-cut dolomicrite and forming isopachous cements lining voids and bed-parallel lenses; and (3) late-stage white calcite spar filling voids and veins cross-cutting ^{13}C-depleted calcite (Fig. 1; Supplementary Figs 2–6).

Application of carbonate clumped isotope thermometry to these materials provides new constraints on their origin and diagenetic history. This technique measures the degree of 'clumping' of heavy isotopes of carbon and oxygen (^{13}C and ^{18}O, respectively) in the carbonate lattice in comparison with a random distribution. The degree of clumping, expressed as Δ_{47} in units of ‰, shows a systematic dependence on temperature[13]. This thermometer can be used to reconstruct the temperature of carbonate precipitation and the oxygen isotope composition of the fluids ($\delta^{18}O_{SMOW}$) from which analysed carbonates formed.

The cap samples examined in this study have Δ_{47} values ranging from 0.487‰ to 0.265‰, corresponding to temperatures of 86–476 °C (Figs 1 and 2a; Supplementary Table 1). In all samples, dolomicrite records the lowest temperatures (mean, 112 °C; $n = 4$), whereas the strongly ^{13}C-depleted grey calcite has the highest temperatures (mean, 378 °C; $n = 4$), with white calcite spar showing intermediate temperatures (mean, 156 °C; $n = 3$). The possibility that high carbonate clumped isotope temperatures result from nonlinear mixing or kinetic effects has been considered and ruled out (Supplementary Discussion). Further discussion of data from dolomicrite and white calcite spar is provided in the Supplementary Discussion; here we focus on the origin of highly ^{13}C-depleted calcite.

Carbonate clumped isotope temperatures from highly ^{13}C-depleted calcite are up to 200–300 °C higher than in any natural carbonate sample analysed to date. Even marbles that experienced ~500 °C during regional metamorphism have lower apparent carbonate clumped temperatures (~200 °C) because cooling over millions of years allowed C–O bonds to keep rearranging until a final 'blocking' temperature was reached[13]. Preservation of high temperatures in the Doushantuo Formation therefore requires faster cooling rates, which we propose are related to a local, short-lived thermal anomaly. More specifically, we suggest that highly ^{13}C-depleted calcite precipitated from a pulse of hot hydrothermal fluid. The exact dependence of blocking temperature on cooling rate is currently unknown. The only potentially relevant constraint is that rearrangement of oxygen by self-diffusion at 500 °C over length scales similar to a unit cell is likely to occur in minutes[14]. Conversely, carbonate blocking temperatures of ~200 °C show that the rate of oxygen diffusion drops rapidly with decreasing temperature. Therefore, the lowest temperatures measured in highly ^{13}C-depleted calcite (~275 °C) could have persisted over geological timescales.

A series of independent observations are consistent with a post-depositional, hydrothermal origin for the highly ^{13}C-depleted calcite. First, by combining temperatures with $\delta^{18}O_{PDB}$ values of carbonate, we calculate that ^{13}C-depleted calcite precipitated from ^{18}O-enriched fluids ($\delta^{18}O_{SMOW} \geq +18‰$), which were distinct from the fluids that precipitated dolomicrite and calcite spar ($\delta^{18}O_{SMOW}$ of $1 \pm 1‰$ and $13 \pm 2‰$, respectively; Fig. 2b, Supplementary Table 1). ^{18}O-enrichment of fluids in seep environments can be caused by clathrate dissociation[7,15]. However, the maximum oxygen isotope fractionation during clathrate formation is +3.2‰ relative to the water source[16]; therefore clathrate dissociation cannot account for the degree of $\delta^{18}O_{SMOW}$ enrichment calculated for fluids precipitating ^{13}C-depleted calcites. Conversely, high-$\delta^{18}O_{SMOW}$ fluids (similar to those that precipitated ^{13}C-depleted calcite) are reported from other ancient continental hydrothermal systems where ^{18}O-rich carbonate host rocks control oxygen isotope fluid compositions[17]. Therefore, we think that the high $\delta^{18}O_{PDB}$ values observed in ^{13}C-depleted calcite (Fig. 2c) are not a sign of exceptional preservation (as previously suggested, ref. 5), but instead are a result of precipitation from hot fluids buffered by ^{18}O-enriched carbonate host rocks (i.e. at low water/rock ratios).

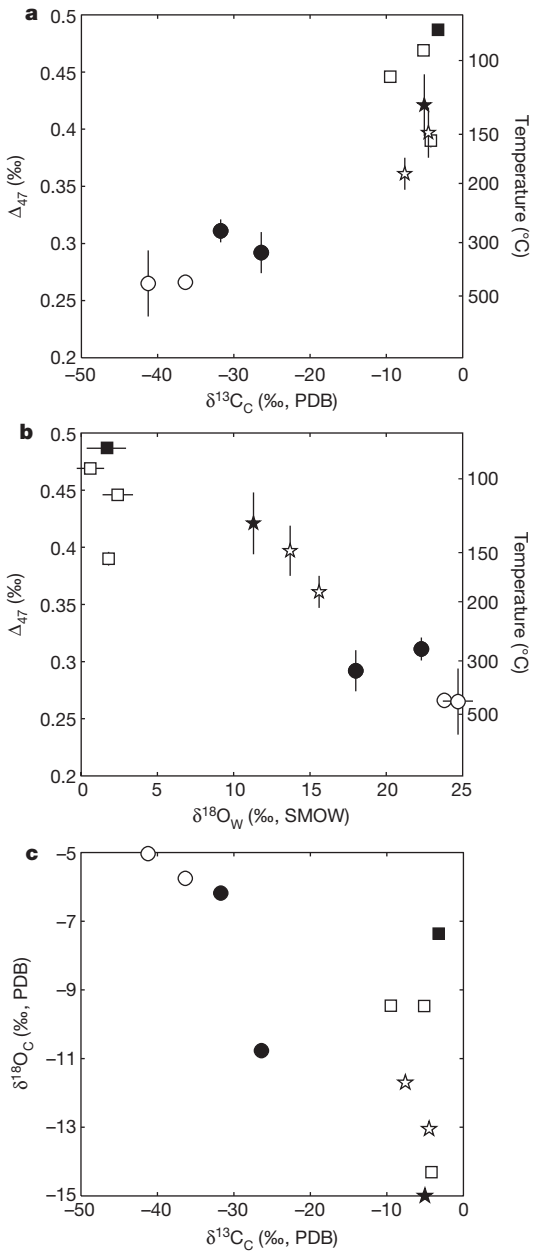


Figure 2 | Cross-plots of Δ_{47} and conventional stable isotope data from the three carbonate phases identified in cap dolostone samples. **a**, Mean Δ_{47} (left axis) and derived temperature (right axis) versus $\delta^{13}C_{PDB}$ of carbonate ($\delta^{13}C_C$). **b**, Mean Δ_{47} (left axis) and derived temperature (right axis) versus the calculated mean $\delta^{18}O_{SMOW}$ of water precipitating carbonates ($\delta^{18}O_W$). **c**, Mean $\delta^{18}O_{PDB}$ of carbonate ($\delta^{18}O_C$) versus $\delta^{13}C_{PDB}$ of carbonate. Error bars, ±1 s.d. based on two or more replicate analyses of the same powder. Error bars are less than the width and height of the symbol where not apparent. Open symbols, samples from the Huajipo section; filled symbols, samples from Jiulongwan. Stars, white calcite spar; squares, dolomicrite; and circles, ^{13}C-depleted grey calcite.

Second, elemental analysis shows that highly ^{13}C-depleted calcite has Mn/Sr ratios >100 (Supplementary Table 2). These cements also have $^{87}Sr/^{86}Sr$ values (0.7090 to 0.7130) that are radiogenic in comparison with the best-preserved Marinoan cap dolostones (0.7072 to 0.7080)[18] and the low Mn/Sr (<1) carbonates from argillaceous dolomites overlying the Doushantuo Formation cap (~0.7080, ref. 19; Fig. 3, Supplementary Table 3). In contrast, most well preserved Phanerozoic seep carbonates have $^{87}Sr/^{86}Sr$ values close to contemporaneous sea water and Mn/Sr ratios <1 (Supplementary Tables 4 and 5). However, it is noteworthy that dolomicrite and white calcite spar also have relatively high Mn/Sr and $^{87}Sr/^{86}Sr$ ratios (Supplementary Table 2), indicating they were diagenetically altered[20]. This raises the possibility that the Mn/Sr and $^{87}Sr/^{86}Sr$ ratios of ^{13}C-depleted calcite may not be uniquely diagnostic of its origin (Supplementary Discussion). And it also makes conceivable the preservation of depositional $\delta^{13}C_{PDB}$ values in highly ^{13}C-depleted calcite, if Δ_{47} were to be reset during exchange of isotopes and trace elements between pre-existing carbonate and hot, carbon-poor fluid with high $\delta^{18}O$, high Mn and high $^{87}Sr/^{86}Sr$. But such a process strikes us as unlikely, because it should have affected other phases in the cap dolostone rather than being exclusive to highly ^{13}C-depleted calcite.

Highly ^{13}C-depleted diagenetic carbonates (with $\delta^{13}C_{PDB}$ values down to −41‰), interpreted as a product of thermogenic oxidation of low-$\delta^{13}C_{PDB}$ hydrocarbons such as methane, have been reported from several other sedimentary basins[21,22]. Similarly, we infer that the extremely low-$\delta^{13}C_{PDB}$ calcite in Doushantuo formed via thermochemical oxidation of hydrothermal methane. We think methane was sourced from organic-rich marls of the Doushantuo Formation, as there are no older source rocks in the Yangtze Gorges area. This is an appealing hypothesis, because the most ^{13}C-depleted methane (down to −51‰, PDB) previously observed in continental hydrothermal systems is produced by the thermogenic breakdown of organic matter in host sediments[23]. In addition, we suggest that the low permeability of overlying clay-rich lithologies of the Doushantuo Formation acted as a seal, causing preferential hydrothermal fluid flow through the porous cap dolostone. This explains why highly ^{13}C-depleted carbonates have only been reported from the cap.

Systematic variations in the degree of clay mineral diagenesis indicate a localized thermal anomaly at the base of the Doushantuo Formation, thus supporting the hypothesis of focused hydrothermal flow. The main clay mineral in the lower 80 m of the Doushantuo Formation in this area is saponite, an Mg-rich trioctahedral smectite interpreted as forming at the time of deposition, based on the unusual mineralogy[24]. However, X-ray diffraction (XRD) data[24] and a decrease in cation exchange capacity (CEC) normalized to the total clay content (Fig. 3; Supplementary Discussion) show that saponite is increasingly altered to corrensite (ordered mixed-layer trioctahedral smectite/chlorite) and chlorite down-section, as the cap dolostone is approached. Because the extent of chloritization of saponite during diagenesis increases with temperature and duration of thermal activity, and does not require extensive fluid exchange, it is a useful means of monitoring maximum palaeotemperatures and is therefore commonly used to study the diagenetic history of sedimentary basins[25]. At temperatures >300 °C, indicated by carbonate clumped isotopes, chloritization takes place at timescales of hundreds to thousands of years[25], consistent with our hypothesis that low Δ_{47} values in carbonates were preserved by rapid cooling.

The thermal gradient implied by clay minerals in the basal 25 m of the Jiulongwan section (Fig. 3) also requires that hydrothermal activity took place after deposition of at least this much sediment. U–Pb zircon ages of 635.2 ± 0.5 Myr and 632.5 ± 0.6 Myr from ash beds within the cap dolostone and 5 m above the top of the cap, respectively[10], show that hydrothermal activity occurred more than 1.6 Myr after deposition of the cap dolostone and is therefore unrelated to deglaciation. We suspect a much younger age, however, coinciding with a regionally extensive Early Cambrian hydrothermal episode in south China (ref. 26; Supplementary Discussion).

Our findings show that the highly ^{13}C-depleted calcite in the Doushantuo Formation is not a record of clathrate destabilization associated with Marinoan deglaciation. The results also highlight a broader puzzle. Before this study, the Doushantuo Formation was considered to contain the oldest examples of carbonates derived by methane oxidation at a cold seep, despite evidence for metabolism of methane dating back to the Archaean[27] and predictions of higher

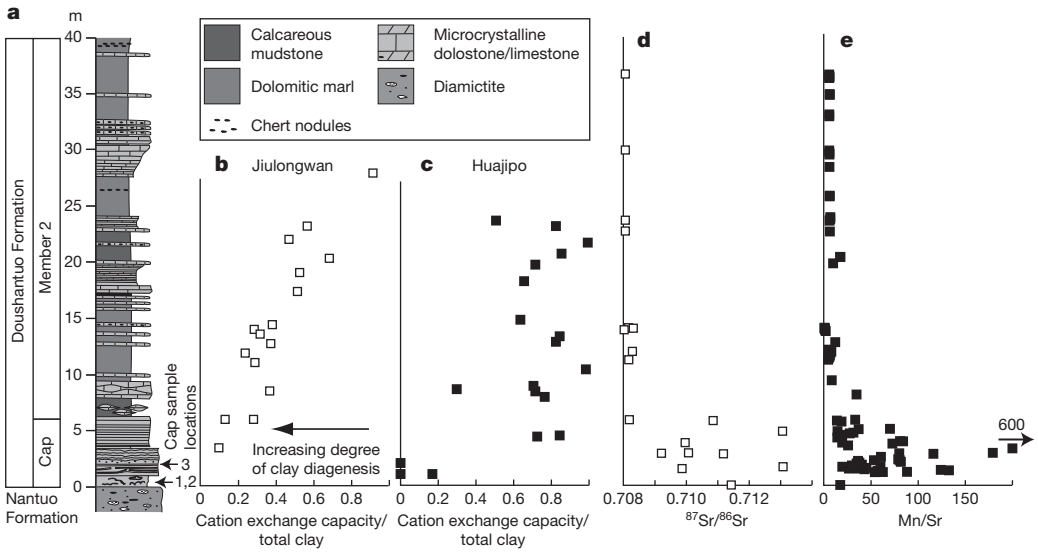


Figure 3 | Stratigraphic variability in trace element content, $^{87}Sr/^{86}Sr$ ratios and degree of chloritization of trioctahedral clay minerals of the lower Doushantuo Formation from the two sections examined in this study. **a**, Stratigraphy. **b**, **c**, Stratigraphic trends in the cation exchange capacity (CEC) of a sample normalized to the total trioctahedral clay content. This ratio is used as an indicator of the extent of clay mineral diagenesis (Supplementary Discussion). Samples with lower normalized CEC contain more corrensite and chlorite and are more diagenetically altered. The stratigraphic position of mineral transformations is different in each section; this is an expected consequence of the heterogeneity of a thermal anomaly induced by hypothesized hydrothermal activity. $^{87}Sr/^{86}Sr$ ratios (**d**) and Mn/Sr data (**e**) are from ref. 19 and are consistent with phase specific trace element and $^{87}Sr/^{86}Sr$ data from the cap dolostone (Supplementary Tables 2 and 3). The sampling locations of cap dolostone samples analysed are shown next to the stratigraphic column. Sample 1 comes from Jiulongwan, samples 2 and 3 come from Huajipo.

fluxes of biogenic methane in the Precambrian[28]. However, our re-interpretation of the highly ^{13}C-depleted calcite, combined with a survey of seep occurrences through Earth history[15] and a recent survey of Precambrian carbon isotope data (including over 11,000 analyses)[29], show that Precambrian carbonates are devoid of cements with $\delta^{13}C_{PDB}$ values less than −30‰, which are characteristic of methane seep carbonates in the Phanerozoic[15].

We propose that this absence reflects the importance of anaerobic oxidation of methane (AOM), using sulphate, in generating exceptionally low $\delta^{13}C_{PDB}$ signatures in carbonate rocks associated with methane seeps. AOM is a biologically mediated reaction that generates ^{13}C-depleted carbonate alkalinity, promoting subsequent precipitation of carbonate with exceptionally low $\delta^{13}C_{PDB}$. Incubation experiments on cold seep sediments, naturally enriched in methanotrophic communities, show that decreased sulphate concentrations result in reduced rates of AOM[30]. Therefore, low sulphate concentrations that characterized Precambrian oceans[9] would have reduced AOM rates, making conditions less favourable for the precipitation of highly ^{13}C-depleted carbonates[11]. Our reinterpretation of the highly ^{13}C-depleted carbonates in the Doushantuo Formation thus highlights the influence of ocean chemistry on methane cycling through Earth history.

METHODS SUMMARY

We characterized the petrography and mineralogy of three cap dolostone samples using visible light microscopy, elemental mapping (with the electron microprobe JEOL JXA 8200 at the California Institute of Technology and an XGT X-ray fluorescence scanner at the Jet Propulsion Laboratory) and quantitative elemental spot analysis of carbonates by electron microprobe. Representative carbonates were micro-drilled from slabs for isotope measurements. Sr isotope measurements were made on acetic acid digestions of carbonate powders using a Neptune MC-ICPMS at the Keck Laboratory, University California, Santa Cruz. Clumped and traditional carbon and oxygen isotope analysis were made on a MAT 253 at the California Institute of Technology. The degree of chloritization of clay minerals was quantified by measuring the CEC to total clay content ratios of 35 samples of the cap and 25 m of overlying sediments collected at Huajipo and Jiulongwan. Quantitative XRD and CEC measurements were carried out at the University of California, Riverside, on whole-rock powders (see Methods and Supplementary Discussion for more details).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 August 2010; accepted 25 March 2011.
Published online 25 May 2011.

1. MacDonald, G. J. Role of methane clathrates in past and future climates. *Clim. Change* **16,** 247–281 (1990).
2. Dickens, G. R., O'Neil, J. R., Rea, D. K. & Owen, R. M. Dissociation of oceanic methane hydrate as a cause of the carbon isotope excursion at the end of the Paleocene. *Paleoceanography* **10,** 965–971 (1995).
3. Hesselbo, S. P. *et al.* Massive dissociation of gas hydrate during a Jurassic oceanic anoxic event. *Nature* **406,** 392–395 (2000).
4. Kennedy, M. J., Christie-Blick, N. & Sohl, L. E. Are Proterozoic cap carbonates and isotopic excursions a record of gas hydrate destabilization following Earth's coldest intervals? *Geology* **29,** 443–446 (2001).
5. Jiang, G. Q., Kennedy, M. J. & Christie-Blick, N. Stable isotopic evidence for methane seeps in Neoproterozoic postglacial cap carbonates. *Nature* **426,** 822–826 (2003).
6. Jiang, G. Q., Kennedy, M. J., Christie-Blick, N., Wu, H. C. & Zhang, S. H. Stratigraphy, sedimentary structures, and textures of the late Neoproterozoic Doushantuo cap carbonate in south China. *J. Sedim. Res.* **76,** 978–995 (2006).
7. Kennedy, M., Mrofka, D. & von der Borch, C. Snowball Earth termination by destabilization of equatorial permafrost methane clathrate. *Nature* **453,** 642–645 (2008).
8. Wang, J. S., Jiang, G. Q., Xiao, S. H., Li, Q. & Wei, Q. Carbon isotope evidence for widespread methane seeps in the ca. 635 Ma Doushantuo cap carbonate in south China. *Geology* **36,** 347–350 (2008).
9. Kah, L. C., Lyons, T. W. & Frank, T. D. Low marine sulphate and protracted oxygenation of the Proterozoic biosphere. *Nature* **431,** 834–838 (2004).
10. Condon, D. *et al.* U-Pb ages from the Neoproterozoic Doushantuo Formation, China. *Science* **308,** 95–98 (2005).
11. Hoffman, P. F. & Schrag, D. P. The snowball Earth hypothesis: testing the limits of global change. *Terra Nova* **14,** 129–155 (2002).
12. Grotzinger, J. P. & Knoll, A. H. Anomalous carbonate precipitates; is the Precambrian the key to the Permian? *Palaios* **10,** 578–596 (1995).
13. Ghosh, P. *et al.* ^{13}C–^{18}O bonds in carbonate minerals: a new kind of paleothermometer. *Geochim. Cosmochim. Acta* **70,** 1439–1456 (2006).
14. Farver, J. R. Oxygen self-diffusion in calcite: dependence on temperature and water fugacity. *Earth Planet. Sci. Lett.* **121,** 575–587 (1994).
15. Campbell, K. A. Hydrocarbon seep and hydrothermal vent paleoenvironments and paleontology: past developments and future research directions. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **232,** 362–407 (2006).
16. Maekawa, T. Experimental study on isotopic fractionation in water during gas hydrate formation. *Geochem. J.* **38,** 129–138 (2004).
17. Bechtel, A., Savin, S. M. & Hoernes, S. Oxygen and hydrogen isotopic composition of clay minerals of the Bahloul Formation in the region of the Bou Grine zinc-lead ore deposit (Tunisia): evidence for fluid-rock interaction in the vicinity of salt dome cap rock. *Chem. Geol.* **156,** 191–207 (1999).
18. Halverson, G. P., Dudas, F. O., Maloof, A. C. & Bowring, S. A. Evolution of the $^{87}Sr/^{86}Sr$ composition of Neoproterozoic seawater. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **256,** 103–129 (2007).
19. Sawaki, Y. *et al.* The Ediacaran radiogenic Sr isotope excursion in the Doushantuo Formation in the Three Gorges area, South China. *Precambr. Res.* **176,** 46–64 (2010).
20. Jacobsen, S. B. & Kaufman, A. J. The Sr, C and O isotopic evolution of Neoproterozoic seawater. *Chem. Geol.* **161,** 37–57 (1999).
21. Boles, J. R., Eichhubl, P., Garven, G. & Chen, J. Evolution of a hydrocarbon migration pathway along basin-bounding faults: evidence from fault cement. *Bull. Am. Assoc. Petrol. Geol.* **88,** 947–970 (2004).
22. Machel, H. G., Cavell, P. A. & Patey, K. S. Isotopic evidence for carbonate cementation and recrystallization, and for tectonic expulsion of fluids into the Western Canada Sedimentary Basin. *Geol. Soc. Am. Bull.* **108,** 1108–1119 (1996).
23. Welhan, J. A. Origins of methane in hydrothermal systems. *Chem. Geol.* **71,** 183–198 (1988).
24. Bristow, T. F. *et al.* Mineralogical constraints on the paleoenvironments of the Ediacaran Doushantuo Formation. *Proc. Natl Acad. Sci. USA* **106,** 13190–13195 (2009).
25. Meunier, A. *Clays* (Springer, 2005).
26. Chen, D., Wang, J., Qing, H., Yan, D. & Li, R. Hydrothermal venting activities in the Early Cambrian, South China: petrological, geochronological and stable isotope constraints. *Chem. Geol.* **258,** 168–181 (2009).
27. Eigenbrode, J. L. & Freeman, K. H. Late Archean rise of aerobic microbial ecosystems. *Proc. Natl Acad. Sci. USA* **103,** 15759–15764 (2006).
28. Kasting, J. F. Methane and climate during the Precambrian era. *Precambr. Res.* **137,** 119–129 (2005).
29. Knauth, L. P. & Kennedy, M. J. The late Precambrian greening of the Earth. *Nature* **460,** 728–732 (2009).
30. Wegener, G. & Boetius, A. An experimental study on short-term changes in the anaerobic oxidation of methane in response to varying methane and sulfate fluxes. *Biogeosciences* **6,** 867–876 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank G. Jiang for guidance in the field and for providing samples, V. Orphan and W. Fischer for discussion and advice, M. Kennedy for supporting fieldwork and use of analytical equipment, and E. Peterman and K. Morrison for laboratory work. C. Ma and M. Anderson are thanked for analytical assistance and advice. This work was supported by an O.K. Earl Postdoctoral fellowship (to T.F.B.), by the NSF EAR and GEG programmes (to J.M.E.), and by INSU (to M.B.). Part of the work of M.B. is IPGP contribution 3138.

Author Contributions T.F.B. and J.P.G. conceived the study. M.B. carried out clumped and conventional isotope analysis and wrote part of the Supplementary Discussion. J.M.E. provided laboratory facilities for isotope work. A.D. and T.F.B. carried out clay mineral analysis. T.F.B. carried out petrographic work and trace element analysis and wrote the manuscript. All authors discussed results, planned analyses and contributed to the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to T.F.B. (thomas.f.bristow@nasa.gov).

METHODS

Clumped and traditional isotope measurements. CO_2 was extracted from all carbonate samples by phosphoric acid digestion at 90 °C using the automatic laboratory methods described elsewhere[31]. CO_2 was analysed at the California Institute of Technology using a Finnigan MAT 253 gas source mass spectrometer configured to collect masses 44–49. Each measurement consisted of eight acquisitions, with typical standard deviations of 0.01‰ to 0.04‰ for Δ_{47} measurements. Values of $\delta^{18}O$ and $\delta^{13}C$ were acquired as part of each analysis and typically show standard deviations one order of magnitude lower (averages of 0.008‰ and 0.004‰, respectively). Carbonate powders with known compositions and heated CO_2 standards were run with unknown samples for offset correction and standardization. Heated gas (CO_2 heated for two hours at 1,000 °C to achieve a stochastic isotopic distribution) with a range of bulk stable isotope compositions similar to samples ($\delta^{13}C \ll 0$‰) were analysed to minimize the potential errors associated with mass spectrometric nonlinearities. For this study, data were normalized relative to a fixed heated gas line model (for each of the three sessions of analyses run) and raw data corrected for instrument nonlinearity and scale compression as described in ref. 32. Δ_{47} data were then corrected for an acid reaction temperature of 90 °C, with the correction factor of 0.081‰, as experimentally determined[31]. Finally, Δ_{47} data were normalized to working carbonate standards: a vein calcite named 102-GC-AZ01 and an Italian marble named Carrara marble. This last correction stage is typically less than 0.02‰. Fourteen distinct extractions of Carrara marble standard during this period yielded a mean Δ_{47} value of 0.356 ± 0.012‰, a $\delta^{13}C_{PDB}$ of 2.34 ± 0.04‰ and a $\delta^{18}O_{PDB}$ of -1.75 ± 0.07‰. Five separate extractions of 102-GC-AZ01 yielded a Δ_{47} of 0.646 ± 0.009‰, a $\delta^{13}C_{PDB}$ of 0.51 ± 0.06‰ and a $\delta^{18}O_{PDB}$ of -14.32 ± 0.06‰. Accepted Δ_{47} values for these standards from >60 analyses by multiple analysts in our laboratory are 0.352‰ and 0.654‰ for Carrara marble and 102-GC-AZ01, respectively.

Fractionation factors of 1.00821 and 1.0093 were used to account for the temperature-dependent oxygen isotope fractionation between CO_2 gas and carbonates resulting from the reaction with phosphoric acid at 90 °C, for calcitic samples[33] and dolomitic samples[34], respectively. Measured values of Δ_{47} (in ‰) were used to estimate carbonate growth temperature, using the empirically derived polynomial determined using high temperature experimental carbonates[13], hydrothermal dolomite (M.B. *et al.*, manuscript in preparation) and inorganic synthetic calcites[13]. Paired temperature and carbonate $\delta^{18}O_{PDB}$ data were used to calculate the $\delta^{18}O_{SMOW}$ values of hydrothermal and/or formation waters that interacted with analysed carbonates, using temperature dependent carbonate–water fractionations described in ref. 35 for calcite and ref. 36 for dolomite. All samples were analysed at least twice using sub-fractions of the same powder, to account for heterogeneity. Uncertainties in temperature estimates and isotopic data (Fig. 2; Supplementary Table 1) are based on a minimum of two analyses of the same powder. A minimum error for Δ_{47} measurements of 0.02‰ has been applied based on the maximum external precision expected from counting statistics; see, for example, ref. 32.

Sr ratios. Because of the relatively large amount of sample required for combined Δ_{47} and Sr isotope analysis, powders for Sr isotope analysis of sample 3 were collected from the same phases in a companion slab cut from the sample used to obtain powders for isotope measurements. For other samples, powders were obtained by re-drilling holes initially used to collect powder for Δ_{47} measurements. To avoid potential contamination from silicate phases, Sr isotope measurements were made on acetic acid digests of carbonate powders, using a Neptune MC-ICPMS at the Keck Laboratory, University California, Santa Cruz. The precision of these measurements is 0.00002, based on repeated measurements of an internal standard that yielded a mean Sr ratio of 0.71030 ($n = 5$). The accepted value of this standard is 0.71025 and this offset was applied in correcting measured sample ratios. A repeat measurement of a sample of highly ^{13}C-depleted calcite from sample 3 yielded Sr ratios that were within 0.00002.

Elemental mapping and analysis. Two methods were used in elemental mapping. Areas of samples 1 and 3, outlined in blue in Supplementary Fig. 4 and 5, were scanned using an XGT X-ray fluorescence scanner at the Jet Propulsion Laboratory, Pasadena, California. The images give qualitative information about the elemental abundances in various phases of the samples, with the brightest areas containing the highest relative abundance of a particular element.

A highly polished, large format (51 × 75 mm) thin section, of part of sample 3, corresponding to the area inside the blue box shown in Supplementary Fig. 5, was carbon coated for electron probe analysis using the JEOL JXA 8200 instrument at the Division of Geological and Planetary Sciences Analytical Facility, California Institute of Technology. A representative 6 × 6 mm area that contained each of the three main carbonate phases was selected for elemental mapping. The elemental maps shown in Supplementary Fig. 6 give qualitative information about elemental abundance, with brighter areas containing higher concentrations of an element.

Quantitative measurements of Mg, Ca, Sr, Fe, Mn, Si, Al and Ba in the various carbonate phases were made using multiple spot analyses, with a defocused 10 μm spot at 15 kV and 15 nA.

Quantitative XRD and CEC measurements. These analyses were made at the University of California, Riverside, on powdered samples from the same sections of the Doushantuo Formation (at Huajipo and Jiulongwan) that cap dolostone samples were collected from (Fig. 3). The methods used for analysis are described elsewhere[24,37].

31. Passey, B. H., Levin, N. E., Cerling, T. E., Brown, F. H. & Eiler, J. M. High-temperature environments of human evolution in East Africa based on bond ordering in paleosol carbonates. *Proc. Natl Acad. Sci. USA* **107,** 11245–11249 (2010).
32. Huntington, K. W. *et al.* Methods and limitations of 'clumped' CO_2 isotope (Δ_{47}) analysis by gas-source isotope ratio mass spectrometry. *J. Mass Spectrom.* **44,** 1318–1329 (2009).
33. Swart, P. K., Burns, S. J. & Leder, J. J. Fractionation of the stable isotopes of oxygen and carbon in carbon-dioxide during the reaction of calcite with phosphoric acid as a function of temperature and technique. *Chem. Geol.* **86,** 89–96 (1991).
34. Rosenbaum, J. & Sheppard, S. M. F. An isotopic study of siderites, dolomites and ankerites at high-temperatures. *Geochim. Cosmochim. Acta* **50,** 1147–1150 (1986).
35. Friedman I. & O'Neil J. R.. Compilation of stable isotope fractionation factors of geochemical interest. *Prof. Pap. US Geol. Surv.* **440-KK,** (1977).
36. Horita, J. The dolomite problem: oxygen isotope fractionation to elevated temperatures. *Geochim. Cosmochim. Acta* **72,** A391 (2008).
37. Orsini, L. & Remy, J. C. Utilisation du chlorure de cobaltihexammine pour la détermination simultanée de la capacité d'échange et des bases échangeables des sols. *Sci. Sol* **4,** 269–275 (1976).

LETTER

doi:10.1038/nature10110

Transcriptomic analysis of autistic brain reveals convergent molecular pathology

Irina Voineagu[1], Xinchen Wang[2], Patrick Johnston[3], Jennifer K. Lowe[1], Yuan Tian[1], Steve Horvath[4], Jonathan Mill[3], Rita M. Cantor[4], Benjamin J. Blencowe[2] & Daniel H. Geschwind[1,4]

Autism spectrum disorder (ASD) is a common, highly heritable neurodevelopmental condition characterized by marked genetic heterogeneity[1–3]. Thus, a fundamental question is whether autism represents an aetiologically heterogeneous disorder in which the myriad genetic or environmental risk factors perturb common underlying molecular pathways in the brain[4]. Here, we demonstrate consistent differences in transcriptome organization between autistic and normal brain by gene co-expression network analysis. Remarkably, regional patterns of gene expression that typically distinguish frontal and temporal cortex are significantly attenuated in the ASD brain, suggesting abnormalities in cortical patterning. We further identify discrete modules of co-expressed genes associated with autism: a neuronal module enriched for known autism susceptibility genes, including the neuronal specific splicing factor *A2BP1* (also known as *FOX1*), and a module enriched for immune genes and glial markers. Using high-throughput RNA sequencing we demonstrate dysregulated splicing of *A2BP1*-dependent alternative exons in the ASD brain. Moreover, using a published autism genome-wide association study (GWAS) data set, we show that the neuronal module is enriched for genetically associated variants, providing independent support for the causal involvement of these genes in autism. In contrast, the immune-glial module showed no enrichment for autism GWAS signals, indicating a non-genetic aetiology for this process. Collectively, our results provide strong evidence for convergent molecular abnormalities in ASD, and implicate transcriptional and splicing dysregulation as underlying mechanisms of neuronal dysfunction in this disorder.

We analysed post-mortem brain tissue samples from 19 autism cases and 17 controls from the Autism Tissue Project and the Harvard brain bank (Supplementary Table 1) using Illumina microarrays. For each individual, we profiled three regions previously implicated in autism[5]: superior temporal gyrus (STG, also known as Brodmann's area (BA) 41/42), prefrontal cortex (BA9) and cerebellar vermis. After filtering for high-quality array data (Methods), we retained 58 cortex samples (29 autism, 29 controls) and 21 cerebellum samples (11 autism, 10 controls) for further analysis (see Methods for detailed sample description). We identified 444 genes showing significant expression changes in autism cortex samples (DS1, Fig. 1b), and only 2 genes were differentially expressed between the autism and control groups in cerebellum (Methods), indicating that gene expression changes associated with autism were more pronounced in the cerebral cortex, which became the focus of further analysis (Supplementary Table 2). There was no significant difference in age, post mortem interval (PMI), or RNA integrity numbers (RIN) between autism and control cortex samples (Supplementary Fig. 1, Methods).

Supervised hierarchical clustering based on the top 200 differentially expressed genes showed distinct clustering of the majority of autism cortex samples (Fig. 1a), including one case that was simultaneously found to have a 15q duplication (Methods, Supplementary Table 1), which is known to cause 1% of ASD[6]. Cortex samples from ten of the cases coalesced in a single tight-clustering branch of the dendrogram. Clustering was independent of age, sex, RIN, PMI, co-morbidity of seizures, or medication (Fig. 1a and Supplementary Fig. 2c). It is interesting to note that the two ASD cases that cluster with controls (Fig. 1a) are the least severe cases, as assessed by global functioning (Supplementary Table 12). We observed a highly significant overlap between differentially expressed genes in frontal and temporal cortex ($P = 10^{-44}$; Fig. 1b), supporting the robustness of the data and indicating that the autism-specific expression changes are consistent across these cortical areas. We also validated a cross section of the differentially expressed genes by quantitative reverse transcription PCR (RT–PCR) and confirmed microarray-predicted changes in 83% of the genes tested (Methods, Supplementary Fig. 2b). Gene ontology enrichment analysis (Methods) showed that the 209 genes downregulated in autistic cortex were enriched for gene ontology categories related to synaptic function, whereas the upregulated genes ($N = 235$) showed enrichment for gene ontology categories implicated in immune and inflammatory response (Supplementary Table 3).

To test whether these findings were replicable, and to further validate the results in an independent data set, we obtained tissue from an additional frontal cortex region (BA44/45) from nine ASD cases and five controls (DS2; Supplementary Table 4). Three of the cases and all of the controls used for validation were independent from our initial cohort. Ninety-seven genes were differentially expressed in BA44/45 in DS2, and 81 of these were also differentially expressed in our initial cohort ($P = 1.2 \times 10^{-93}$, hypergeometric test; Fig. 1b, c). Remarkably, the direction of expression differences between autism and controls was the same as in the initial cohort for all but 2 of the 81 overlapping differentially expressed probes. Hierarchical clustering of DS2 samples based on either the top 200 genes differentially expressed in the initial cohort or the 81 overlapping genes showed distinct separation of cases from controls (Supplementary Fig. 6). In addition, comparison of these differentially expressed results with another, smaller study of the STG in ASD[7], revealed significant consistency at the level of differentially expressed genes, including downregulation of *DLX1* and *AHI1* (Supplementary Table 5). Thus, differential expression analysis produced robust and highly reproducible results, warranting further refined analysis.

We next applied weighted-gene co-expression network analysis (WGCNA)[8,9] to integrate the expression differences observed between autistic and control cerebral cortex into a higher order, systems level context. We first asked whether there are global differences in the organization of the brain transcriptome between autistic and control brain by constructing separate co-expression networks for the autism and control groups (Methods). The control brain network showed high similarity with the previously described human brain co-expression networks (Supplementary Table 7), consistent with the existence of

[1]Program in Neurogenetics and Neurobehavioral Genetics, Department of Neurology and Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, California 90095-1769, USA. [2]Banting and Best Department of Medical Research, Donnelly Centre, University of Toronto, Toronto, Ontario M5G 1L6, Canada. [3]Institute of Psychiatry, King's College London, London SE5 8AF, UK. [4]Department of Human Genetics, University of California Los Angeles, Los Angeles, California 90095, USA.

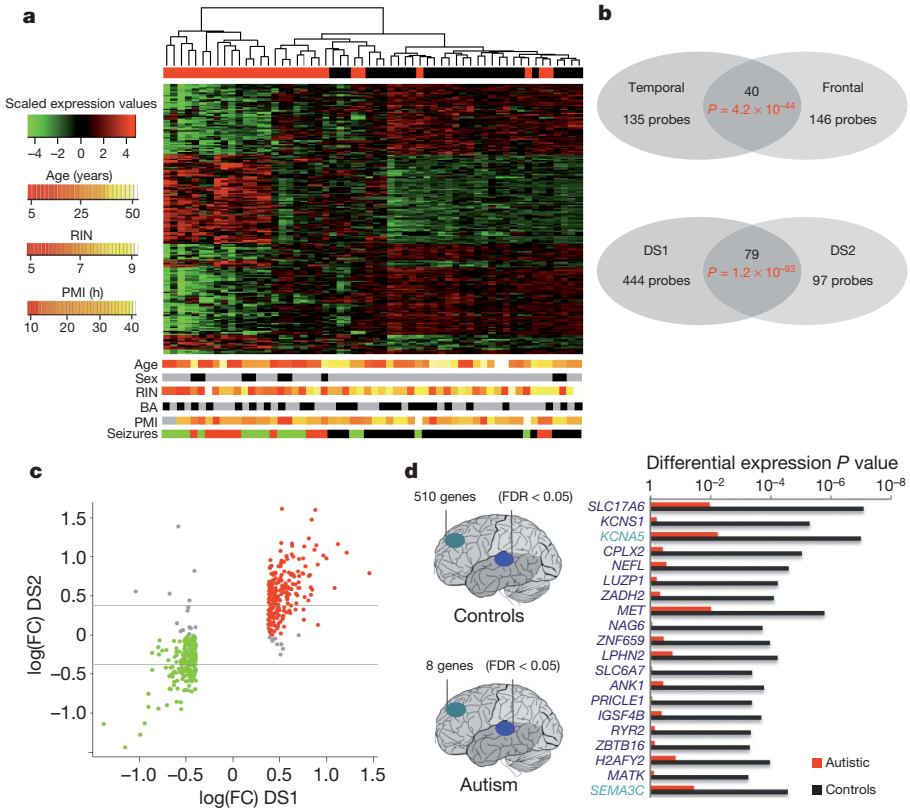


Figure 1 | Gene expression changes in autism cerebral cortex **a**, Heat map of top 200 genes differentially expressed between autism and control cortex samples. Scaled expression values are colour-coded according to the legend on the left. The dendrogram depicts hierarchical clustering based on the top 200 differentially expressed genes. The top bar (A/C) indicates the disease status: red, autism; black, control. The bottom bars show additional variables for each sample: sex (grey, male; black, female), brain area (black, temporal; grey, frontal), co-morbidity of seizures (green, autism case with seizure disorder; red, autism case without seizure disorder; black, control), age, RNA integrity number (RIN) and post mortem interval (PMI). BA, Brodmann's area. The corresponding scale for quantitative variables is shown on the left. **b**, Top, Venn diagram depicting the overlap between genes differentially expressed in frontal and temporal cortex. Bottom, Venn diagram describing the overlap between genes differentially expressed in the initial cohort (DS1) and the replication cohort (DS2). Differential expression in the initial cohort was assessed at an FDR < 0.05 and fold change > 1.3. The statistical criteria were relaxed to $P < 0.05$ for the replication data set because it involved fewer samples. **c**, Expression fold changes for all genes differentially expressed in the initial cohort are plotted on the x-axis against the fold changes for the same genes in the replication cohort on the y-axis. Green, genes downregulated in the autism group in both data sets; red, genes upregulated in the autism group in both data sets; grey, genes with opposite direction of variation in the two data sets. Horizontal lines show fold change threshold for significance. **d**, Diagram depicting the number of genes showing significant expression differences between frontal and temporal cortex in control samples (top) and autism samples (bottom) at FDR < 0.05 (left). The top 20 genes differentially expressed between frontal and temporal cortex in control samples (right). All of the genes shown are also differentially expressed between frontal and temporal cortex in fetal midgestation brain[10], but show no significant expression differences between frontal and temporal cortex in autism. The horizontal bars depict P values for differential expression between frontal and temporal cortex in the autism and control groups.

robust modules of co-expressed genes related to specific cell types and biological functions[8]. Similarly, the majority (87%) of the autism modules showed significant overlap with the previously described human brain modules (Supplementary Table 6), indicating that many features reflecting the general organization of the autism brain transcriptome are consistent with that of the normal human brain.

The expression levels of each module were summarized by the first principal component (the module eigengene), and were used to assess whether modules are related to clinical phenotypes or other experimental variables, such as brain region. Two of the control module eigengenes (cM6, cM13) showed significant differences ($P < 0.05$) between the two cortical regions as expected, whereas none of the ASD modules showed any differences between frontal and temporal cortex. This led us to explore the hypothesis that the normal molecular distinctions between the two cortical regions tested were altered in ASD compared with controls. Remarkably, whereas 174 genes were differentially expressed between control BA9 and BA41 (false discovery rate (FDR) $< 1\%$), none of the genes were differentially expressed in the same regional comparison among the ASD cases. This was not simply an issue of statistical thresholds, as relaxing the statistical criteria for differential expression to an FDR of 5% identified over 500 differentially expressed genes in controls, and only 8 in ASD brains, confirming the large difference observed in regional cortical differential gene expression between ASD cases and controls (Fig. 1d, Methods). Analysis of differential expression from a data set[10] of gene expression in developing fetal human brain showed a highly significant ($P = 5.8 \times 10^{-9}$) overlap of differentially expressed genes with those found in controls in this study, independently confirming that these genes differentiate normal temporal and frontal lobes. We evaluated the homogeneity of gene expression variance across the autism and control groups using Bartlett's test (Methods) which indicated that increased variance was not the major factor responsible for the striking difference in regional gene expression between ASD and controls (Supplementary Fig. 7 and Supplementary Data).

These data suggest that typical regional differences, many of which are observed during fetal development[10], are attenuated in frontal and temporal lobe in autism brain, pointing to abnormal developmental patterning as a potential pathophysiological driver in ASD. This is especially interesting in light of a recent anatomical study of five cases with adult autism which demonstrated a reduction in typical ultrastructural differences between three frontal cortical regions in autism[11]. Together, these independent studies provide both molecular and

structural evidence suggesting a relative diminution of cortical regional identity in autism.

To identify discrete groups of co-expressed genes showing transcriptional differences between autism and controls, we constructed a co-expression network using the entire data set, composed of both autism and control samples (Methods). As previously shown for complex diseases[12,13] co-expression networks allow analysis of gene expression variation related to multiple disease-related and genetic traits. We assessed module eigengene relationship to autism disease status, age, gender, cause of death, co-morbidity of seizures, family history of psychiatric disease, and medication, providing a complementary assessment of these potential confounders to that performed in the standard differential expression analysis (Supplementary Table 9).

The comparison between autism and control groups revealed two network modules whose eigengenes were highly correlated with disease status, and not any of the potential confounding variables (Supplementary Table 9). We found that the top module (M12) showed highly significant enrichment for neuronal markers (Supplementary Table 9), and high overlap with two neuronal modules previously identified as part of the human brain transcriptional network[8]: a *PVALB*+ interneuron module and a module of genes involved in synaptic function. The M12 eigengene was under-expressed in autism cases, indicating that genes in this module were downregulated in the autistic brain (Fig. 2). Consistent with the pathways identified to be down-regulated in autism by differential expression analysis (Supplementary Table 3), the functional enrichment of M12 included the gene ontology categories involved in synaptic function, vesicular transport and neuronal projection.

Remarkably, unlike differentially expressed genes, M12 showed significant overrepresentation of known autism susceptibility genes[2] (Supplementary Table 10; $P = 6.1 \times 10^{-4}$), including *CADPS2*, *AHI1*, *CNTNAP2*, and *SLC25A12*, supporting the increased power of the network-based approach to identify disease-relevant transcriptional changes. A further advantage of network analysis over standard analysis of differential expression is that it allows one to infer the functional relevance of genes based on their network position[9]. The hubs of M12, that is, the genes with the highest rank of M12 membership[8], were *A2BP1*, *APBA2*, *SCAMP5*, *CNTNAP1*, *KLC2*, and *CHRM1* (Supplementary Data). The first three of these genes have previously been implicated in autism[14–16], whereas the fourth is a homologue of

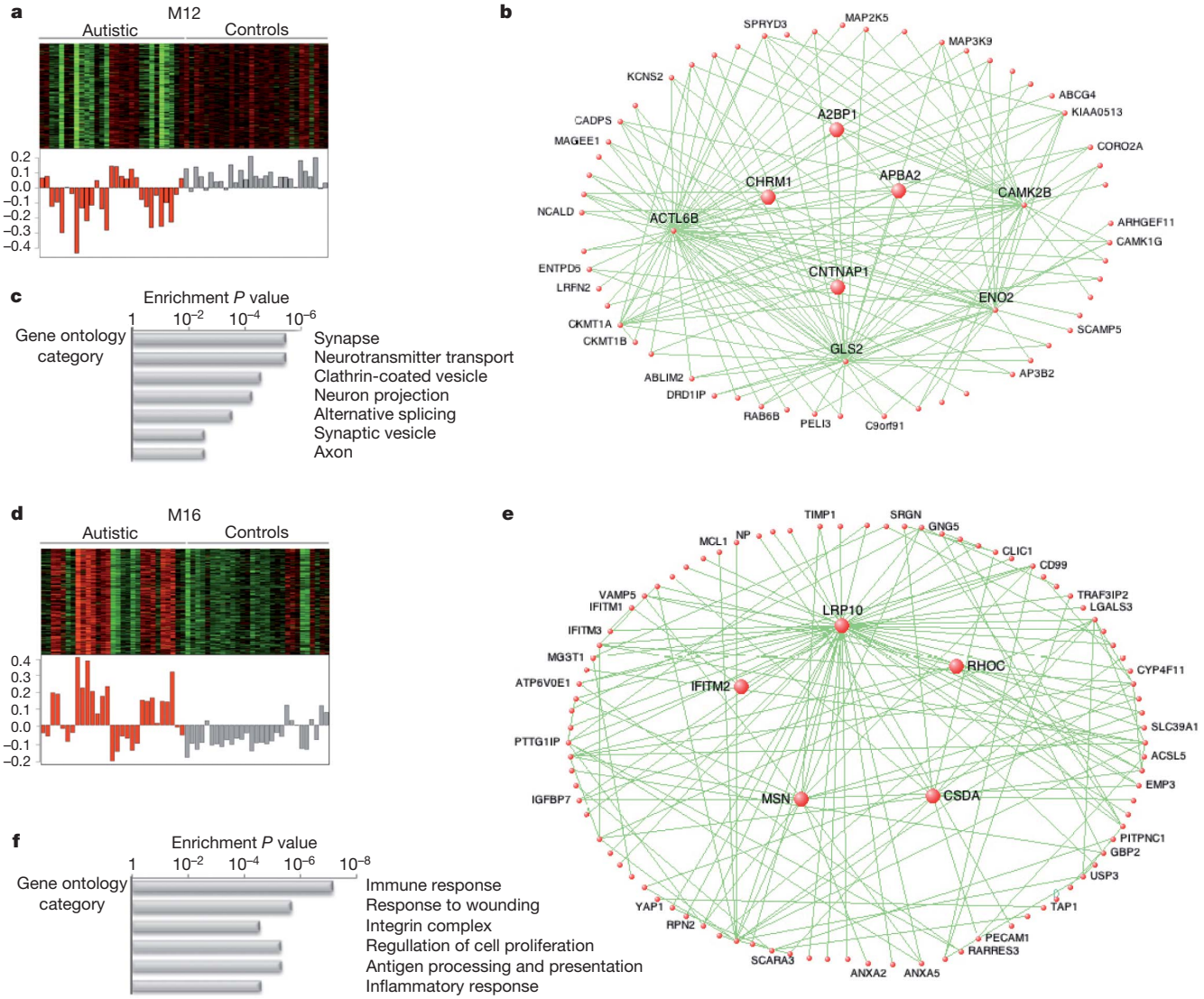


Figure 2 | Gene co-expression modules associated with autism **a**, **d**, Heat map of genes belonging to the co-expression module (top). Corresponding module eigengene values (*y*-axis) across samples (*x*-axis) (bottom). Red, autism; grey, controls. **b**, **e**, Visualization of the M12 and M16 modules, respectively. The top 150 connections are shown for each module. Genes with the highest correlation with the module eigengene value (that is, intramodular hubs) are shown in larger size. **c**, **f**, Relevant gene ontology categories enriched in the M12 and M16 modules.

the autism susceptibility gene *CNTNAP2* (ref. 17). We highlight the group of genes most strongly connected to the known ASD genes (Supplementary Fig. 5) and emphasize the downregulation of several interneuron markers, such as *DLX1* and *PVALB*, as candidates for future genetic and pathologic investigations.

The second module of co-expressed genes highly related to autism disease status, M16, was enriched for astrocyte markers and markers of activated microglia (Supplementary Table 9), as well as for genes belonging to immune and inflammatory gene ontology categories (Fig. 2). This module, which was upregulated in ASD brain, showed significant similarity to two modules identified in previous studies of normal human brain[8]: an astrocyte module and a microglial module. Consistent with this functional annotation, two of the hubs of the M16 module were known astrocyte markers (*ADFP*, also known as *PLIN2*, and *IFITM2*).

One of the hubs of the M12 module was *A2BP1*, a neural- and muscle-specific alternative splicing regulator[18] and the only splicing factor previously implicated in ASD[16]. Because *A2BP1* was downregulated in several ASD cases (Supplementary Fig. 8), this observation provided a unique opportunity to identify potential disease-relevant *A2BP1* targets. Whereas *A2BP1*-regulated alternative exons have been predicted genome-wide[19], few genes have been experimentally validated as *A2BP1* targets[20]. To identify potential *A2BP1*-dependent differential splicing events in ASD brain, we performed high-throughput RNA sequencing (RNA-Seq) on three autism samples with significant downregulation of *A2BP1* (average fold change by quantitative RT–PCR = 5.9) and three control samples with average *A2BP1* levels. We identified 212 significant alternative splicing events (Supplementary Data). Among these, 36 had been defined[19] as predicted targets of *A2BP1/2*, which represents a highly significant overlap (36/176; $P = 2.2 \times 10^{-16}$). In addition, five previously validated *A2BP1* targets showed evidence of alternative splicing, four of which (*ATP5C1*, *ATP2B1*, *GRIN1* and *MEF2C*) were confirmed as having differential splicing between ASD samples with low *A2BP1* expression and control samples, indicating that we were able to identify a high proportion of the expected *A2BP1*-dependent differential splicing events. We also observe that alternative exons with increased skipping in ASD relative to control cases are significantly enriched for *A2BP1* motifs in adjacent, downstream intronic sequences ($P = 1.09 \times 10^{-7}$, Fisher's exact test), consistent with previous data[19].

The top gene ontology categories enriched among ASD differential splicing genes highly overlapped with the gene ontology categories found to be enriched in the M12 module (Fig. 3b). In addition, *A2BP1* target genes showed enrichment for actin-binding proteins and genes involved in cytoskeleton reorganization (Fig. 3b). Among top predicted *A2BP1*-dependent differential splicing events (Fig. 3a) are *CAMK2G*, which also belongs to the M12 module, as well as *NRCAM* and *GRIN1*. The latter are proteins involved in synaptogenesis, in which allelic variants have been associated with autism and schizophrenia, respectively[21,22].

RT–PCR assays confirmed a high proportion (85%) of the tested differential splicing changes involving predicted A2BP1 targets (Supplementary Fig. 8). We further tested the differential splicing events validated by RT–PCR in three independent ASD cases with decreased *A2BP1* levels and confirmed the predicted changes in alternative splicing (Supplementary Fig. 8), indicating that the observed differential splicing events are indeed associated with reduced *A2BP1* levels, rather than due to inter-individual variability. The RNA-Seq data thus provides validation of the functional groups of genes identified by co-expression analysis, and evidence for a convergence of transcriptional and alternative-splicing abnormalities in the synaptic and signalling pathogenesis of ASD.

To test whether our findings are more generalizable, and determine whether the autism-associated transcriptional differences observed are likely to be causal, versus collateral effects or environmentally-induced changes, we tested whether our co-expression modules or

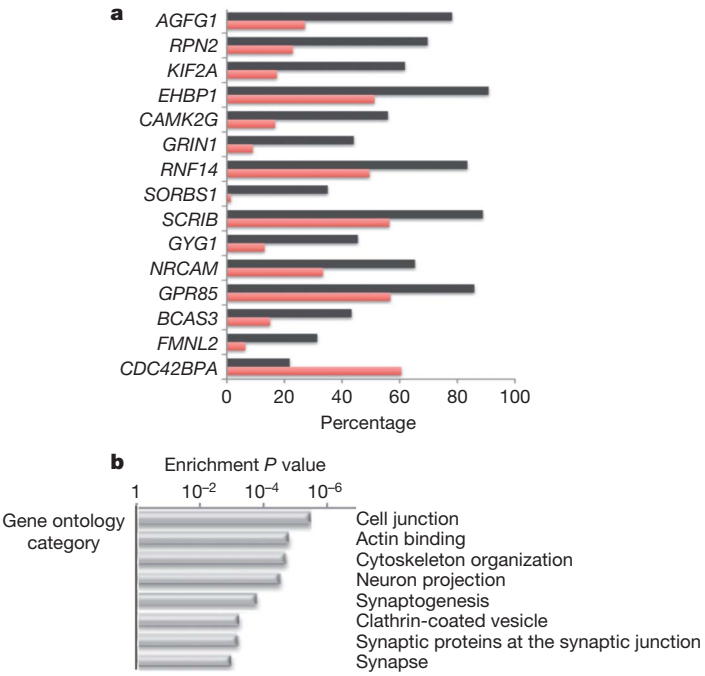


Figure 3 | A2BP1-dependent differential splicing events **a**, Top A2BP1-specific differential splicing events. Differential splicing events showing the most significant differences in alternative splicing between low-A2BP1 autism cases and controls as well as differential splicing differences consistent with the A2BP1 binding site position. The horizontal axis depicts the percentage of transcripts including the alternative exon. Red, autism samples; black, control samples. **b**, Relevant gene ontology categories enriched in the set of genes containing exons differentially spliced between low-A2BP1 autism cases and controls.

the differentially expressed genes show enrichment for autism genetic association signals. M12 showed highly significant enrichment for association signals ($P = 5 \times 10^{-4}$), but neither M16 nor the list of differentially expressed genes showed such enrichment (Fig. 4). As a negative control, we performed the same set-enrichment analysis using two GWAS studies for non-psychiatric disease performed on the same genotyping platform: a genome-wide association for hair colour[23], and a GWAS study of warfarin maintenance dose[24] finding no significant enrichment of the association signal (Fig. 4b, Supplementary Fig. 4). These results indicate that (1) M12 consists of a set of genes that are supported by independent lines of evidence to be causally involved in ASD pathophysiology, and (2) the upregulation of immune response genes in the autistic brain observed by us and others[25] has no evidence of a common genetic component.

Our system-level analysis of the ASD brain transcriptome demonstrates the existence of convergent molecular abnormalities in ASD for the first time, providing a molecular neuropathological basis for the disease, whose genetic, epigenetic, or environmental aetiologies can now be directly explored. The genome-wide analysis performed here significantly extends previous findings implicating synaptic dysfunction, as well as microglial and immune dysregulation in ASD[6] by providing an unbiased systematic assessment of transcriptional alterations and their genetic basis. We show that the transcriptome changes observed in ASD brain converge with GWAS data in supporting the genetic basis of synaptic and neuronal signalling dysfunction in ASD, whereas immune changes have a less pronounced genetic component and thus are most likely either secondary phenomena or caused by environmental factors. Because immune molecules and cells such as microglia have a role in synaptic development and function[26], we speculate that the observed immune upregulation may be related to abnormal ongoing plasticity in the ASD brain. The striking attenuation of gene expression differences observed here between frontal and temporal cortex in ASD is likely to represent a defect of developmental patterning and provides a strong rationale for further studies to assess the pervasiveness

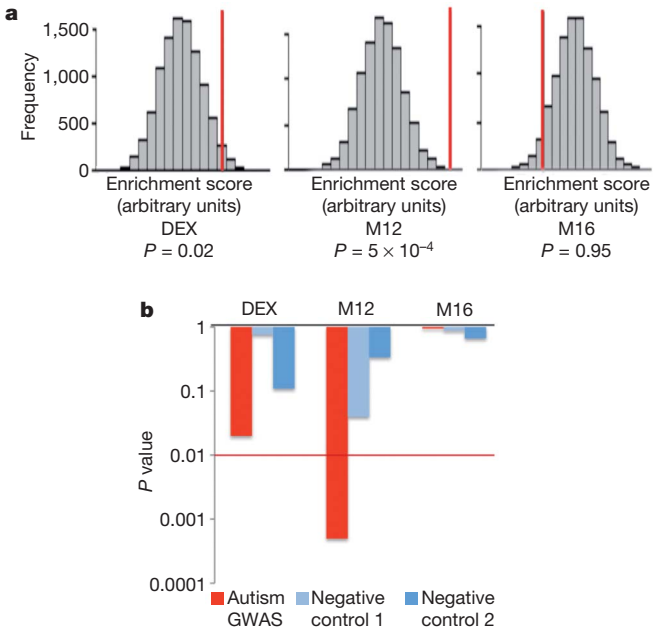


Figure 4 | GWAS set enrichment analysis **a**, GWAS set enrichment analysis using the discovery AGRE cohort from ref. 27. For each gene set (DEX, differentially expressed genes; M12 and M16) the null distribution of the enrichment score generated by 10,000 random permutations is shown (*x*-axis) and the enrichment score for the gene set is depicted by a red vertical line. A *P* value <0.01 was considered significant to correct for multiple comparisons. **b**, GWAS signal enrichment of differentially expressed genes and the autism-associated co-expression modules M12 and M16. Enrichment *P* values are shown for an autism GWAS data set (ref. 27, AGRE discovery cohort) as well as two control data sets consisting of GWAS studies of non-psychiatric traits: ref. 23 (Negative control 1) and ref. 24 (Negative control 2). The red line marks the *P* value threshold for significance.

of transcriptional patterning abnormalities across the ASD brain. We also demonstrate for the first time alterations in differential splicing associated with *A2BP1* levels in the ASD brain, and show that many of the affected exons belong to genes involved in synaptic function. Finally, given current evidence of genetic overlap between ASD and other neurodevelopmental disorders including schizophrenia and attention deficit hyperactivity disorder (ADHD), the data provide a new pathway-based framework from which to assess the enrichment of genetic association signals in other allied psychiatric disorders.

METHODS SUMMARY

Brain tissue. Post-mortem brain tissue was obtained from the Autism Tissue Project and the Harvard Brain Bank as well as the MRC London Brain bank for Neurodegenerative Disease. Detailed information on the autism cases included in this study is available in Methods.

Microarrays and RNA-seq. Total RNA was extracted from 100 mg of tissue using a Qiagen miRNA kit according to the manufacturer's protocol. Expression profiles were obtained using Illumina Ref8 v3 microarrays. RNA-seq was performed on the Illumina GAIIx, as per the manufacturer's instructions. Further detailed information on data analysis is available in Methods.

Full detailed Methods accompany this paper as Supplementary Information.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 12 December 2010; accepted 13 April 2011.
Published online 25 May 2011.

1. Durand, C. M. *et al.* Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. *Nature Genet.* **39,** 25–27 (2006).
2. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466,** 368–372 (2010).
3. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316,** 445–449 (2007).
4. Geschwind, D. H. Autism: many genes, common pathways? *Cell* **135,** 391–395 (2008).
5. Amaral, D. G., Schumann, C. M. & Nordahl, C. W. Neuroanatomy of autism. *Trends Neurosci.* **31,** 137–145 (2008).
6. Abrahams, B. S. & Geschwind, D. H. Advances in autism genetics: on the threshold of a new neurobiology. *Nature Rev. Genet.* **9,** 341–355 (2008).
7. Garbett, K. *et al.* Immune transcriptome alterations in the temporal cortex of subjects with autism. *Neurobiol. Dis.* **30,** 303–311 (2008).
8. Oldham, M. C. *et al.* Functional organization of the transcriptome in human brain. *Nature Neurosci.* **11,** 1271–1282 (2008).
9. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4,** 17 (2005).
10. Johnson, M. B. *et al.* Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* **62,** 494–509 (2009).
11. Zikopoulos, B. & Barbas, H. Changes in prefrontal axons may disrupt the network in autism. *J. Neurosci.* **30,** 14595–14609 (2010).
12. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452,** 429–435 (2008).
13. Plaisier, C. L. *et al.* A systems genetics approach implicates *USF1*, *FADS3*, and other causal candidate genes for familial combined hyperlipidemia. *PLoS Genet.* **5,** e1000642 (2009).
14. Babatz, T. D., Kumar, R. A., Sudi, J., Dobyns, W. B. & Christian, S. L. Copy number and sequence variants implicate *APBA2* as an autism candidate gene. *Autism Res.* **2,** 359–364 (2009).
15. Castermans, D. *et al.* *SCAMP5*, *NBEA* and *AMISYN*: three candidate genes for autism involved in secretion of large dense-core vesicles. *Hum. Mol. Genet.* **19,** 1368–1378 (2010).
16. Martin, C. L. *et al.* Cytogenetic and molecular characterization of *A2BP1/FOX1* as a candidate gene for autism. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **144B,** 869–876 (2007).
17. Alarcón, M. *et al.* Linkage, association, and gene-expression analyses identify *CNTNAP2* as an autism-susceptibility gene. *Am. J. Hum. Genet.* **82,** 150–159 (2008).
18. Underwood, J. G., Boutz, P. L., Dougherty, J. D., Stoilov, P. & Black, D. L. Homologues of the *Caenorhabditis elegans* Fox-1 protein are neuronal splicing regulators in mammals. *Mol. Cell. Biol.* **25,** 10005–10016 (2005).
19. Zhang, C. *et al.* Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. *Genes Dev.* **22,** 2550–2563 (2008).
20. Lee, J. A., Tang, Z. Z. & Black, D. L. An inducible change in Fox-1/A2BP1 splicing modulates the alternative splicing of downstream neuronal target exons. *Genes Dev.* **23,** 2284–2293 (2009).
21. Moy, S. S., Nonneman, R. J., Young, N. B., Demyanenko, G. P. & Maness, P. F. Impaired sociability and cognitive function in *Nrcam*-null mice. *Behav. Brain Res.* **205,** 123–131 (2009).
22. Zhao, X. *et al.* Significant association between the genetic variations in the 5′ end of the N-methyl-D-aspartate receptor subunit gene *GRIN1* and schizophrenia. *Biol. Psychiatry* **59,** 747–753 (2006).
23. Han, J. *et al.* A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genet.* **4,** e1000074 (2008).
24. Cooper, G. M. *et al.* A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood* **112,** 1022–1027 (2008).
25. Morgan, J. T. *et al.* Microglial activation and increased microglial density observed in the dorsolateral prefrontal cortex in autism. *Biol. Psychiatry* **68,** 368–376 (2010).
26. Boulanger, L. M. Immune proteins in brain development and synaptic plasticity. *Neuron* **64,** 93–109 (2009).
27. Wang, K. *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* **459,** 528–533 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful for the efforts of the Autism Tissue Program (ATP) of Autism Speaks and the families that have enrolled in the ATP, which made this work possible. We also thank S. Scherer and R. Wintle for sharing their SNP genotyping data on the AGP samples with us before its publication. We would also like to thank B. Abrahams for help in the initial stages of the project, B. Fogel, G. Konopka, N. Barbosa-Morais and J. Bomar for critically reading the manuscript, M. Lazaro for help with tissue dissection, and C. Vijayendran and K. Winden for useful discussions. This work was funded by an Autism Center of Excellence Network Grant from NIMH 5R01MH081754-03 and NIMH R37MH060233 to D.H.G. and by grants from the Canadian Institutes of Health Research and Genome Canada through the Ontario Genomics Institute to B.J.B. and others.

Author Contributions I.V. and D.H.G. designed the study and wrote the manuscript. I.V. performed experiments, analysed the data and conducted the GWAS set enrichment analysis. X.W. and B.J.B. analysed the RNA sequencing data. J.K.L. contributed to the GWAS set enrichment analysis. Y.T. performed some of the microarray qRT-PCR validation experiments. R.M.C. supervised the GWAS set enrichment analysis. S.H. supervised the WGCNA analysis. P.J. and J.M. provided dissected tissue for the replication experiment. All authors discussed the results and commented on the manuscript.

Author Information All microarray and RNA-seq data are deposited in GEO under accession number GSE28521. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.H.G. (dhg@mednet.ucla.edu).

METHODS

Brain tissue samples. Brain tissue samples from 19 autism cases and 17 controls were obtained from the Autism Tissue Project (ATP) and the Harvard Brain Bank. For each brain, tissue was obtained from frontal cortex (BA9), temporal cortex (BA41/42 or BA22) and cerebellum (vermis), with the exception of three controls lacking the cerebellum sample (Supplementary Table 1). For the replication experiment, frontal cortex tissue (BA44/45) from nine ASD cases and five controls were obtained from the ATP and MRC London Brain bank for Neurodegenerative Disease respectively (Supplementary Table 4).

For all of the autism cases, clinical information is available upon request from ATP (http://www.autismtissueprogram.org), including the ADI-R diagnostic scores. Supplementary Table 12 contains a summary of clinical characteristics. Although autism cases with known genetic causes were not included in this study, one case with a chromosome 15q duplication was identified for AN17138 by high density small nucleotide polymorphism (SNP) arrays[28] during the course of this study. The ATP cases were genotyped with high-density SNP arrays and with two exceptions all are Caucasians. The two Asian samples cluster with the other ASD cases in the current study, and are not distinguishable from the Caucasian cases based on clustering by gene expression.

RNA extractions and microarrays. Total RNA was extracted from approximately 100 mg of frozen tissue, using the Qiagen miRNA kit. RNA concentration was assessed by a NanoDrop spectrophotometer and RNA quality was measured using an Agilent Bioanalyzer. All RNA samples included in the expression analysis had an RNA integrity number (RIN) > 5. cDNA labelling and hybridizations on Illumina Ref8 v3 microarrays were performed according to the manufacturer's protocol.

Microarray data analysis. Microarray data analysis was performed using the R software and Bioconductor packages. Raw expression data were $\log_2$ transformed and normalized by quantile normalization. Data quality control criteria included high inter-array correlation (Pearson correlation coefficients > 0.85) and detection of outlier arrays based on mean inter-array correlation and hierarchical clustering. Probes were considered robustly expressed if the detection P value was <0.05 for at least half of the samples in the data set. Cortex samples (58: 29 autism, 29 controls) and cerebellum samples (21: 11 autism, 10 controls) fulfilled all data quality control criteria. The 29 autism cortex samples included tissue from 13 ASD cases with both frontal and temporal cortex and 3 ASD cases with frontal cortex only (in total 16 frontal cortex and 13 temporal cortex ASD samples). The 29 autism control samples also included tissue from 13 controls with both frontal and temporal cortex and 3 controls with frontal cortex only (in total 16 frontal cortex and 13 temporal cortex control samples).

Initially, all samples were normalized together to assess clustering by brain region. As expected, we observed distinct clustering of cortex and cerebellum samples (Supplementary Fig. 2A). For subsequent analyses, cortex samples and cerebellum samples were normalized and analysed separately.

Differential expression. Differential expression was assessed using the SAM package (significance analysis of microarrays, http://www-stat.stanford.edu/~tibs/SAM) and unless otherwise specified the significance threshold was FDR < 0.05 and fold changes > 1.3. Given that SAM is less sensitive in detecting differentially expressed genes for small number of samples, for the replication cohort, the differential expression was assessed by a linear regression method (Limma package, http://bioconductor.org/packages/release/bioc/html/limma.html). Our results showing high degree of overlap between genes differentially expressed in the two data sets indicate that the expression differences observed are independent of the analysis methods.

Because 444 genes were differentially expressed between autism and controls in cortex and only 2 genes were differentially expressed between the two groups in cerebellum (FDR < 0.05), we tested whether this difference was due to the smaller number of cerebellum samples, by relaxing the statistical criteria to FDR < 0.25. We found fewer than 10 differentially expressed genes in cerebellum using the relaxed statistical criteria, supporting the conclusion that genome-wide expression changes in autism were more pronounced in cerebral cortex than in cerebellum.

To account for the fact that the control group of DS1 contained samples from a single female whereas the autism DS1 group included four females, we eliminated from differential expression analysis all probes showing evidence of gender-specific gene expression ($n = 70$). We also applied linear regression of expression values against age and sex, and then assessed differential expression between the autism and control groups using the residual values. We observed a 96% overlap between differentially expressed genes using either the residual values or the raw data, indicating that neither age nor sex were major drivers of expression differences between the autism and control groups.

Differential expression between frontal and temporal cortex was assessed by a paired modified t-test (SAM) using the 13 autism and 13 control cases for which RNA samples from both cortex areas passed the quality control criteria. For each of the 510 genes that were differentially expressed in control samples between frontal and temporal cortex, we compared the variance of autism and control expression values in frontal cortex and temporal cortex. The homogeneity of variance (homoscedasticity) of gene expression was assessed using the Barlett test in R. Fifty one genes showed a significant difference in variance ($P < 0.05$, Barlett test) between autism and control groups both in frontal and temporal cortex, and the Barlett test P-values for these genes are listed in Supplementary Data.

WGCNA. Unsigned co-expression networks were built using the WGCNA package in R. Probes with evidence of robust expression (9,914; see above) were included in the network. Network construction was performed using the blockwiseModules function in the WGCNA package[29], which allows the network construction for the entire data set. For each set of genes a pair-wise correlation matrix is computed, and an adjacency matrix is calculated by raising the correlation matrix to a power. The power of 10 was chosen using the scale-free topology criterion[9] and was used for all three networks: the network built using autism samples only, controls samples only or all samples. An advantage of weighted correlation networks is the fact that the results are highly robust with respect to the choice of the power parameter. For each pair of genes, a robust measure of network interconnectedness (topological overlap measure) was calculated based on the adjacency matrix. The topological overlap based dissimilarity was then used as input for average linkage hierarchical clustering. Finally, modules were defined as branches of the resulting clustering tree. To cut the branches, we used the hybrid dynamic tree-cutting because it leads to robustly defined modules[31]. To obtain moderately large and distinct modules, we set the minimum module size to 40 genes and the minimum height for merging modules at 0.1. Each module was summarized by the first principal component of the scaled (standardized) module expression profiles. Thus, the module eigengene explains the maximum amount of variation of the module expression levels. For each module, we defined the module membership measure (also known as module eigengene based connectivity kME) as the correlation between gene expression values and the module eigengene. Genes were assigned to a module if they had a high module membership to the module (kME > 0.7). An advantage of this definition (and the kME measure) is that it allows genes to be part of more than one module. Genes that did not fulfil these criteria for any of the modules are assigned to the grey module. For the cell type marker enrichment analysis we used the markers defined experimentally in refs 32 and 33 which were previously used to annotate human brain network modules[34,35].

Module visualization: the topological overlap measure was calculated for the top 100 genes in each module ranked by kME. The resulting list of gene pairs was filtered so that both genes in a pair had the highest kME for the module plotted (that is, most module-specific interactions). The resulting top 150 gene pairs were plotted using Visant.

Gene ontology analyses. Functional enrichment was assessed using the DAVID database http://david.abcc.ncifcrf.gov/. For differentially expressed genes and co-expression modules, the background was set to the total list of genes expressed in the brain in the cortex data set. For genes containing differentially spliced exons, the background was set to the total set of genes showing evidence of alternative splicing in our RNA-seq data. The statistical significance threshold level for all gene ontology enrichment analyses was $P < 0.05$ (Benjamini and Hochberg corrected for multiple comparisons).

Statistical analyses. All gene set overlap analyses were performed by assessing the cumulative hypergeometric probability using the phyper function in R. The population size was defined as the total number of probes expressed in both data sets. If the comparison involved different platforms, the comparison was done at gene level.

Quantitative RT–PCR. One microgram of total RNA was treated with RNase-free DNase I (Invitrogen/Fermentas) and reverse-transcribed using Invitrogen Superscript II reverse-transcriptase and random hexanucleotide primers (Invitrogen). Real time PCR was performed on an ABI7900 cycler in 10 µl volume containing iTaq Sybrgreen (Biorad) and primers at a concentration of 0.5 µM each. The results shown in Supplementary Fig. 2b represent at least two independent cDNA synthesis experiments for each gene. *GAPDH* levels were used as an internal control. Statistical significance was assessed by a two-tailed t-test assuming unequal variance.

Semi-quantitative RT–PCR. Total RNA (600 ng) pooled from autism cases ($n = 2$–3) or controls ($n = 2$–3) was reverse-transcribed as described above. cDNA (50 ng) was subjected to 30 cycles of PCR amplification using the primers described in Supplementary Table 11. PCR products were separated on a 3% agarose gel stained with GelStar (Lonza).

RNA sequencing and data analysis.. 73-nucleotide reads were generated using an Illumina GAII sequencer according to the manufacturer's protocol. To generate sufficient read coverage for the quantitative analysis of alternative splicing events, reads for ASD and control brain samples were separately pooled and aligned to an existing database of EST and cDNA-derived alternative splicing junctions using

the Basic Local Alignment Tool (BLAT) as described previously[36,37]. Reads were considered properly aligned to a splice junction if at least 71 of the 73 nucleotides matched and at least 5 nucleotides mapped to each of the two exons forming the splice junction. Alternative exon inclusion values ('%inc'), representing the proportion of messenger RNA transcripts with the alternatively spliced exon included, were calculated for each mRNA pool as the ratio of reads aligning to the C1-A or A-C2 junctions against reads aligning against all three possible junctions as previously described[36] (C1-A, A-C2, C1-C2, see Supplementary Fig. 3). Calculated %inc values were considered reliable if at least one of the included junctions as well as the skipped junctions were covered by at least 20 reads. %inc values were compared across samples using Fisher's exact test and the Bonferroni–Hochberg correction to identify differentially spliced exons associated with autism. Differential splicing events were considered significant if they fulfilled both criteria of FDR < 0.1 and %inc difference between autism and controls $>15\%$.

GWAS set enrichment analysis. GWAS enrichment analysis was performed as previously described in ref. 38 with the main modification that we generated the null distribution, using permutation of gene labels rather than permutation of case/control labels, because the raw genotyping data was not available for all data sets. This approach has been proposed as an acceptable alternative to phenotype label permutation[38] and has been previously used for set enrichment analyses of GWAS data[39]. For all genes that met the robust expression criteria in our data set, we mapped the SNPs present on the Illumina 550k platform located within the transcript boundaries and an additional 20 kb on the 5′ end and 10 kb on the 3′ end. Each gene was assigned a GWAS significance value consisting of the lowest P value of all SNPs mapped to it. A gene set enrichment score (ES) based on the Kolmogorov–Smirnov statistic was calculated as previously described[38] using the $-\log(P\text{-value})$. The null distribution was generated by 10,000 random permutations of gene labels in the list of genes/P-value pairs and an enrichment score ESp was calculated for each permutation. To correct for the gene set size, the enrichment scores were scaled by subtracting the mean and dividing by the standard deviation of ESp. The resulting z-scores were used to calculate the significance p value.

28. Wintle, R. F. *et al.* (2010). A genotype resource for postmortem brain samples from the Autism Tissue Program. *Autism Res.* **4,** 89–97 (2011).
29. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).
31. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24,** 719–720 (2008).
32. Cahoy, J. D. *et al.* A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J. Neurosci.* **28,** 264–278 (2008).
33. Albright, A. V. & Gonzalez-Scarano, F. Microarray analysis of activated mixed glial (microglia) and monocyte-derived macrophage gene expression. *J. Neuroimmunol.* **157,** 27–38 (2004).
34. Oldham, M. C. *et al.* Functional organization of the transcriptome in human brain. *Nature Neurosci.* **11,** 1271–1282 (2008).
35. Miller, J. A., Horvath, S. & Geschwind, D. H. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proc. Natl Acad. Sci. USA* **107,** 12698–12703 (2010).
36. Luco, R. F. *et al.* Regulation of alternative splicing by histone modifications. *Science* **327,** 996–1000 (2010).
37. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genet.* **40,** 1413–1415 (2008).
38. Wang, K., Li, M. & Bucan, M. Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* **81,** 1278–1283 (2007).
39. Zhang, K., Cui, S., Chang, S., Zhang, L. & Wang, J. *i*-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. *Nucleic Acids Res.* **38** (suppl. 2), W90–W95 (2010).

LETTER

doi:10.1038/nature10111

A nuclear-receptor-dependent phosphatidylcholine pathway with antidiabetic effects

Jae Man Lee[1], Yoon Kwang Lee[2,3], Jennifer L. Mamrosh[2], Scott A. Busby[4], Patrick R. Griffin[4], Manish C. Pathak[5], Eric A. Ortlund[5] & David D. Moore[1,2]

Nuclear hormone receptors regulate diverse metabolic pathways and the orphan nuclear receptor LRH-1 (also known as NR5A2) regulates bile acid biosynthesis[1,2]. Structural studies have identified phospholipids as potential LRH-1 ligands[3–5], but their functional relevance is unclear. Here we show that an unusual phosphatidylcholine species with two saturated 12 carbon fatty acid acyl side chains (dilauroyl phosphatidylcholine (DLPC)) is an LRH-1 agonist ligand *in vitro*. DLPC treatment induces bile acid biosynthetic enzymes in mouse liver, increases bile acid levels, and lowers hepatic triglycerides and serum glucose. DLPC treatment also decreases hepatic steatosis and improves glucose homeostasis in two mouse models of insulin resistance. Both the antidiabetic and lipotropic effects are lost in liver-specific *Lrh-1* knockouts. These findings identify an LRH-1 dependent phosphatidylcholine signalling pathway that regulates bile acid metabolism and glucose homeostasis.

Increased fat accumulation in the liver—steatosis—is tightly correlated with insulin resistance and type 2 diabetes[6]. Modestly raised bile acid levels decrease steatosis[7]. Loss of the nuclear receptor LRH-1 decreases bile acid levels[1,2], indicating that an LRH-1 agonist could increase them and improve fatty liver. In screens of a number of different phosphatidylcholine (PC) and other phospholipid species for effects on human LRH-1 transactivation, dilauroyl PC (DLPC; C12:0/C12:0) and diundecanoyl PC (DUPC; C11:0/C11:0) showed strong stimulation (Fig. 1a). Comparable responses were not observed with closely related PCs differing in acyl chain length by only a single methylene group, or with any other C12:0/C12:0 phospholipid species (Supplementary Fig. 1a–c).

DLPC and DUPC, but not the bile acid chenodeoxycholic acid (CDCA) or the more conventional phospholipid dipalmitoyl PC (DPPC; C16:0/C16:0), also activated the synthetic LRH-1 reporter in several other cell lines, including CV-1 and HEK293T cells (data not shown), and specifically increased basal LRH-1 transactivation of the native mouse SHP promoter[8] by approximately twofold in HeLa cells (Supplementary Fig. 2a). DLPC and DUPC also induced a similar response with the OCT4 promoter, which was dependent on both LRH-1 cotransfection and an intact LRH-1 response element[9] (Supplementary Fig. 2a). DLPC and DUPC responsiveness was not altered in mutant LRH-1 derivatives previously shown to inactivate responses to LRH-1 phosphorylation[10] or sumoylation[11], but was strongly decreased by mutations shown to block phospholipid binding[4] (Supplementary Fig. 2d).

Mouse and human LRH-1 showed essentially equivalent responses to DLPC and DUPC, and both DLPC and DUPC also activate the close LRH-1 relative SF-1 (also known as NR5A1; Supplementary Fig. 2). The LRH-1 responses were dose dependent (Supplementary Fig. 2c). Neither DUPC nor DLPC showed significant activation of any of a number of additional nuclear receptors outside of the NR5A subgroup (Supplementary Fig. 2b). In particular, DLPC and DUPC failed to activate PPARα, which was recently reported to be specifically bound and activated by 1-palmitoyl-2-oleoyl (C16:0/C18:1) PC[12], and C16:0/C18:1 PC failed to affect LRH-1 transactivation (Supplementary Fig. 1a). DLPC rapidly induced expression of the LRH-1 target CYP8B1 in the C3A derivative of HepG2 cells (Supplementary Fig. 3a). This response as well as CDCA repression of CYP8B1 expression and transactivation of a synthetic LRH-1 reporter plasmid was specifically compromised in cells transfected with LRH-1 short interfering RNA (siRNA; Supplementary Fig. 3b, c).

We used the mammalian two-hybrid assay and a simple GST pull-down approach to initially test the predicted function of DLPC and DUPC as LRH-1 agonist ligands. In the mammalian two-hybrid analysis, interaction of a VP16–human LRH-1 ligand-binding-domain fusion with a second fusion of the Gal4 DNA-binding domain to the nuclear receptor interaction domain of the coactivator SRC-3 (also known as NCOA3) was unaffected by vehicle, CDCA or DPPC, but was stimulated by either DUPC or DLPC (Supplementary Fig. 4a). *In vitro*, SRC-3 protein did not bind to GST alone but showed a significant basal interaction with a GST–LRH-1-ligand-binding-domain fusion protein, as expected[4]. DLPC and DUPC further increased binding of the coactivator by approximately 3 fold, but vehicle, CDCA, or any of a number of other PC species, including DPPC, had little or no effect (Supplementary Fig. 4b). DLPC also unexpectedly but specifically decreased binding of an SRC-2 peptide to the LRH-1 ligand-binding domain with a half-maximum inhibitory concentration (IC_{50}) of approximately 500 nM, but DPPC had no effect (Supplementary Fig. 4c), and DLPC did not affect rosiglitazone binding to PPARγ (Supplementary Fig. 4d).

As a stringent test of specific binding, the purified bacterially expressed human LRH-1 ligand-binding domain was incubated with DLPC or DPPC at molar ratios of 1:1 or 1:5 (protein:PC), or with buffer alone, and the protein was then repurified to eliminate unbound lipids. Specifically bound lipids were extracted and compared to DLPC or DPPC by electrospray ionization mass spectrometry. Phosphatidylethanolamine (PE) and phosphatidylglycerol (PG) species with 16–22 carbon acyl chain lengths occupy the ligand-binding pocket in the buffer-treated control, with the most abundant peak corresponding to 16:1/18:1 PG (Fig. 1b). DLPC completely replaced these *Escherichia coli* phospholipids, even at an added lipid to protein molar ratio of only 1:1, but DPPC showed no detectable displacement, even at a ratio of 1:5 (Fig. 1b). On the basis of these functional and *in vitro* biochemical results, as well as the extensive structural studies demonstrating phospholipid binding to NR5A receptors[3–5,13,14], we conclude that DLPC and DUPC act *in vitro* as LRH-1 agonists. The functional results indicate that they may also act directly as agonists *in vivo*, although it remains unclear how they might transit the cell membrane and cytosol and enter the nucleus.

PCs are normal dietary nutrients that are efficiently absorbed in the small intestine, and we used the simple route of oral gavage to deliver cholic acid (CA), DLPC, DUPC and DPPC to C5BL/6 mice. These treatments had no apparent toxic effects and did not alter normalized

[1]Program in Developmental Biology, Baylor College of Medicine, Houston, Texas 77030, USA. [2]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, Texas 77030, USA. [3]Department of Integrative Medical Sciences, Northeastern Ohio Universities Colleges of Medicine and Pharmacy, Rootstown, Ohio 44272, USA. [4]The Scripps Research Molecular Screening Center, The Scripps Research Institute, Scripps Florida, Jupiter, Florida 33458, USA. [5]Department of Biochemistry, Emory University School of Medicine, Atlanta, Georgia 30322, USA.

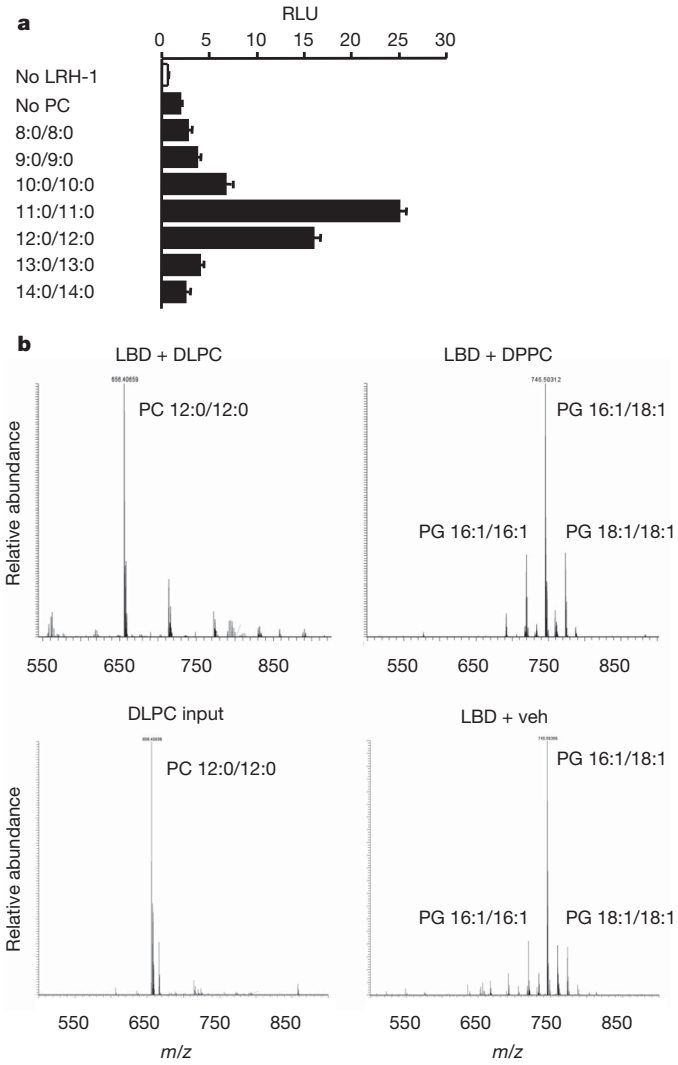


Figure 1 | DLPC activates and binds human LRH-1. a, HeLa cells were transfected with a human LRH-1 expression vector and a luciferase reporter and treated with 100 μM of indicated PCs. Error bars represent mean ± s.e.m. RLU, relative luciferase units. **b**, The human LRH-1 ligand-binding domain (LBD) was expressed and purified as described previously[5] and was incubated at molar ratios of 1:1 and 1:5 (human LRH-1 LBD:PC) with DLPC, DPPC or vehicle (veh) for two hours at 37 °C, and then repurified by size exclusion chromatography to remove unbound phospholipids. Bound lipids were analysed using electrospray mass injection mass spectrometry in the negative-ion mode. Results with DLPC (1:1), DPPC (1:5) and vehicle are shown, along with analysis of re-extracted DLPC; DLPC (1:5) and DPPC (1:1) incubations were very similar to those shown. The re-extracted DPPC peak is at 768.5, and is not detectable in any of the DPPC incubations.

liver weight (Supplementary Fig. 5a) or increase serum indicators of liver damage (Supplementary Fig. 5b). CA reduced expression of CYP7A1 and CYP8B1 and induced SHP, as expected, and DPPC was without significant effect (Fig. 2a). Both DLPC and DUPC significantly induced expression of CYP7A1, CYP8B1 and SR-B1, and repressed SHP (Fig. 2a). The substantial induction of CYP8B1, particularly by DLPC, is in accord with the opposite response in liver-specific *Lrh-1* knockouts[1,2], which otherwise show relatively limited alterations in gene expression or liver physiology. The decreased SHP expression is consistent with the induction of the bile acid biosynthetic enzymes, but was not expected based on the acute response of the isolated SHP promoter in HeLa cells (Supplementary Fig. 2a). Because SHP represses its own expression in the liver[8], it is possible that an initial inductive response is followed by an autoregulatory decrease. The induction of CYP7A1 and CYP8B1 is lost in liver-specific *Lrh-1* knockouts generated by infecting LRH-1 floxed (f/f) mice[2] with adenoviral Cre (Ad-Cre) expression vectors (Supplementary Fig. 6a, b).

These gene expression changes were associated with a modest but significant increase in the total bile acid pool and serum bile acid levels in DLPC- and DUPC-treated mice (Fig. 2b), consistent with the opposite effect in liver-specific *Lrh-1* knockouts[1,2]. These DLPC and DUPC effects were lost in Ad-Cre-mediated liver-specific knockouts (Supplementary Fig. 6c). Both CA- and DUPC/DLPC-treated mice showed significantly decreased serum non-esterized fatty acids (NEFAs) (Fig. 2b) and hepatic triglycerides (Fig. 2c), which were associated with decreased serum glucose in DUPC/DLPC-treated, but not CA-treated mice (Fig. 2b). Serum and hepatic cholesterol levels were unaffected (Fig. 2b, c). As anticipated based on the effective clearance of phospholipid-containing gut-derived chylomicrons by the liver, neither DLPC nor DUPC altered expression of SF-1 target genes in the adrenal gland (Supplementary Fig. 5c).

Prompted by the lipid and glucose effects in the normal mice, we focused on DLPC, a natural product, and treated insulin-resistant leptin-receptor-deficient *db/db* mice for 2 weeks by oral gavage, followed by a glucose tolerance test (GTT). An insulin tolerance test (ITT) was carried out after 1 week of additional treatment. Glucose homeostasis was improved in DLPC-treated mice, as shown by the GTT and ITT (Supplementary Fig. 7a), as well as lower fasting serum insulin levels (Supplementary Fig. 7e). DLPC treatment did not affect body weight or a number of other parameters, but decreased expression of the lipogenic transcription factor SREBP-1c (also known as SREBF1) and its downstream targets, and significantly lowered hepatic triglyceride levels in the *db/db* mice (Supplementary Figs 7b–f and 8).

To critically test the role of hepatic LRH-1 in these antidiabetic effects, wild-type *Lrh-1*$^{f/f}$ and liver-specific *Lrh-1*$^{-/-}$ mice generated using an albumin-Cre transgene (Supplementary Fig. 9a) were fed a high-fat diet to induce obesity and insulin resistance (diet-induced obesity (DIO)) for 15 weeks. Continuing on the diet, they were treated daily by oral gavage with vehicle or DLPC for 3 weeks, and glucose homeostasis was assessed by GTT and ITT. Loss of LRH-1 did not affect glucose homeostasis in the *Lrh-1*$^{-/-}$ DIO mice relative to the *Lrh-1*$^{f/f}$ DIO mice (Fig. 3a). As in the *db/db* mice, DLPC treatment substantially improved glucose homeostasis in the *Lrh-1*$^{f/f}$ DIO mice as indicated by GTT and ITT, and these responses were absent in the *Lrh-1*$^{-/-}$ DIO mice (Fig. 3a). The DLPC treated *Lrh-1*$^{f/f}$ DIO mice also had decreased fasting serum glucose and insulin levels, resulting in an 80% decrease in the homeostatic model assessment of insulin resistance (HOMA-IR) (Fig. 3b). Increased insulin sensitivity was confirmed using the hyperinsulinaemic–euglycaemic clamp, which showed both increased glucose disposal and markedly decreased hepatic glucose production in the DLPC-treated mice (Fig. 3c). Increased overall insulin sensitivity was also confirmed by increased insulin-dependent phosphorylation of the insulin receptor, IRS2 and AKT in the DLPC-treated *Lrh-1*$^{f/f}$ livers, but not *Lrh-1*$^{-/-}$ livers (Fig. 3d and Supplementary Fig. 9c).

Total body weight and food intake (Supplementary Fig. 9b), as well as weights of liver, reproductive fat pads, or brown fat did not differ between the *Lrh-1*$^{f/f}$ and *Lrh-1*$^{-/-}$ DIO mice. However, the livers of DLPC-treated *Lrh-1*$^{f/f}$ DIO mice were less pale and fatty, and decreased lipid deposition was confirmed both histologically and by direct measurement of hepatic triglyceride levels (Fig. 4a, b). NEFA levels were also decreased by DLPC in *Lrh-1*$^{f/f}$, but not *Lrh-1*$^{-/-}$ DIO livers and serum (Fig. 4b). Hepatic and serum bile acid levels were significantly increased by DLPC in the *Lrh-1*$^{f/f}$, but not the *Lrh-1*$^{-/-}$ DIO mice (Fig. 4b). DLPC significantly induced both CYP7A1 and CYP8B1 expression in the *Lrh-1*$^{f/f}$ mice, and this specific response was absent in the *Lrh-1*$^{-/-}$ mice (Supplementary Fig. 10a). The expression of additional bile-acid-related genes, including the biosynthetic CYP7B1 and CYP27A1 and the hepatic bile acid transporters BSEP (also known as ABCB11) and NTCP (also known as SLC10A1) was not significantly affected by DLPC treatment in *Lrh-1*$^{f/f}$ DIO mice (Supplementary Fig. 10a).

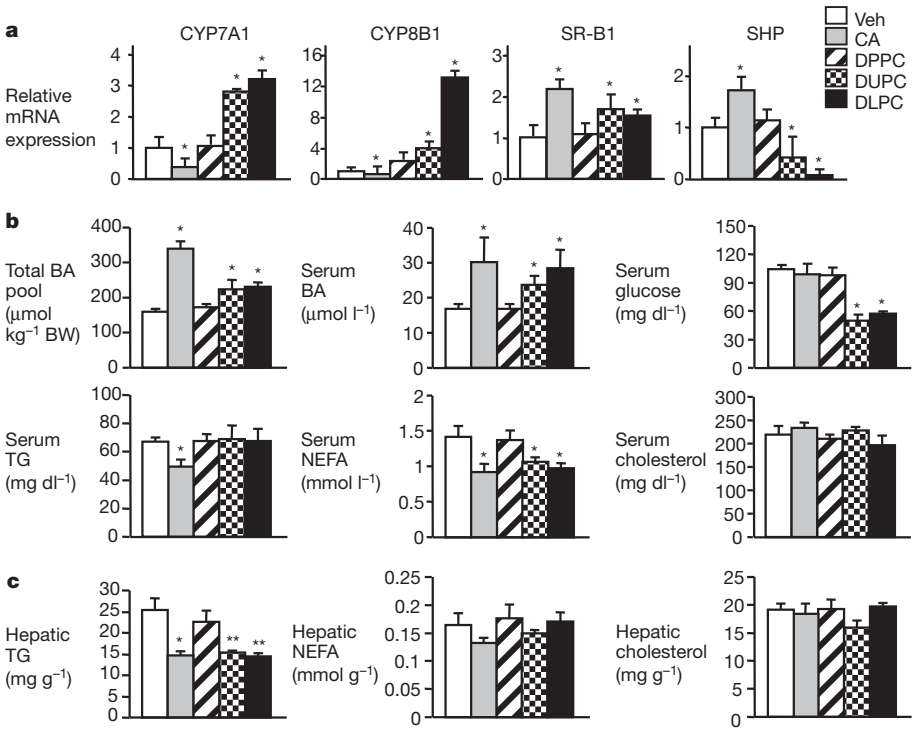


Figure 2 | DLPC and DUPC modulate expression of LRH-1 target genes in liver. a, Eight-week-old male C57BL/6 mice were challenged orally with vehicle (Veh), CA, DPPC, DUPC and DLPC for 3 days. Total liver RNA was isolated and prepared for the complementary DNA. Hepatic gene expression was determined using quantitative polymerase chain reaction (PCR). mRNA levels are relative to 36B4. **b**, Total bile acid (BA) pool and serum BA, glucose, triglyceride (TG), NEFA and cholesterol were measured in the same mice. **c**, Hepatic TG, NEFA and cholesterol were measured in the same mice. Error bars represent mean ± s.e.m. $*P < 0.05$, $**P < 0.01$ versus vehicle; $n = 5$ mice per group.

In accord with *db/db* results (Supplementary Fig. 8), there was little or no effect on hepatic expression of a number of glucose homeostasis and fatty acid oxidation genes (Supplementary Fig. 10d, e). However, DLPC markedly decreased expression of genes associated with *de novo* lipogenesis (Fig. 4c), including the lipogenic transcription factor SREBP-1c and its key downstream targets ACC-2, SCD-1 and FASN in *Lrh-1*$^{f/f}$ DIO mice (Fig. 4c). The beneficial effects of DLPC on glucose homeostasis and fatty liver in *Lrh-1*$^{f/f}$ mice fed a high-fat diet and infected with a control Ad–GFP vector were also lost in mice in which the *Lrh-1*$^{f/f}$ allele was deleted by Ad-Cre expression (Supplementary Fig. 11). Overall, we conclude that LRH-1 is required for the antidiabetic effects of DLPC. However, it remains to be determined whether its effects are a consequence of being a direct ligand for LRH-1, and it remains possible that DLPC activates an alternative signalling cascade or induces biosynthesis of an endogenous LRH-1 ligand.

Here we have identified DLPC as a specific agonist ligand for LRH-1 *in vitro*. Further studies will be needed to address the intriguing questions of whether phospholipid transfer proteins[15] facilitate its transport to the large and dynamic intranuclear pool of phosphatidylcholine[16], and whether DLPC is an endogenous LRH-1 agonist. The ligand responsiveness of LRH-1 is consistent with the identification of synthetic agonists that activate both LRH-1 and SF-1 (ref. 17). When expressed in an adrenal cell line, SF-1 is bound by a relatively low molecular weight form of phosphatidic acid with two saturated 14 carbon acyl chains, which acts as an agonist for SF-1 but not LRH-1 (ref. 18). Earlier results identified the sphingolipids sphingosine and lyso-sphingomyelin as potential endogenous antagonists of SF-1 transactivation[19]. DLPC does not activate other nuclear receptors, including PPARα or PPARγ, which have previously been reported to be activated by more conventional longer chain phospholipid species[12,20,21]. In the opposite direction, LRH-1 is not activated by conventional PC species, including the C16:0/C18:1 PC reported to specifically bind and activate PPARα in the liver[12]. Phospholipids are emerging as a structurally diverse class of highly specific nuclear receptor ligands.

The beneficial effect of DLPC on steatosis is associated with significantly decreased expression of the transcription factor SREBP-1c and its downstream lipogenic targets. At least two complementary mechanisms could contribute to this decrease. As SREBP-1c autoregulates its own expression[22], the reported functional antagonism of SREBP-1c transactivation by LRH-1 (ref. 23) could directly inhibit SREBP-1c promoter activity. As SREBP-1c expression is induced by insulin[24], the DLPC-dependent decrease in serum insulin should also decrease SREBP-1c messenger RNA. The combination of these two mechanisms could set up a positive regulatory loop in which the initial LRH-1-dependent repression of SREBP-1c expression would decrease steatosis and increase insulin sensitivity, resulting in a decrease in serum insulin. This decrease would then reinforce the decline in SREBP-1c expression and activity, further ameliorating fatty liver and thereby continuing a beneficial cycle (Supplementary Fig. 12). This essentially reverses the lipogenic vicious cycle to insulin resistance proposed previously by McGarry[25], and supported by more recent results with SREBP-1c[26].

These beneficial effects are probably complemented by an increase in fatty acid β-oxidation due to the decrease in acetyl-CoA carboxylase-2 (ACC-2) and its product, malonyl-CoA, which allosterically inhibits CPT-1a enzymatic activity and mitochondrial fatty acid uptake[27]. Decreasing ACC-2 activity in response to either specific antisense oligonucleotides[28] or activation of the nuclear receptor CAR[29] increases β-oxidation and has beneficial effects on both steatosis and insulin resistance. Because SCD-1 ablation also protects against hepatic steatosis by decreasing lipogenesis and increasing β-oxidation[30], reduced SCD-1 expression may also increase β-oxidation in response to LRH-1 activation.

We conclude that the identification of DLPC as a useful tool for analysis of LRH-1 function has uncovered an unexpected, LRH-1-dependent PC signalling pathway that can improve fatty acid and

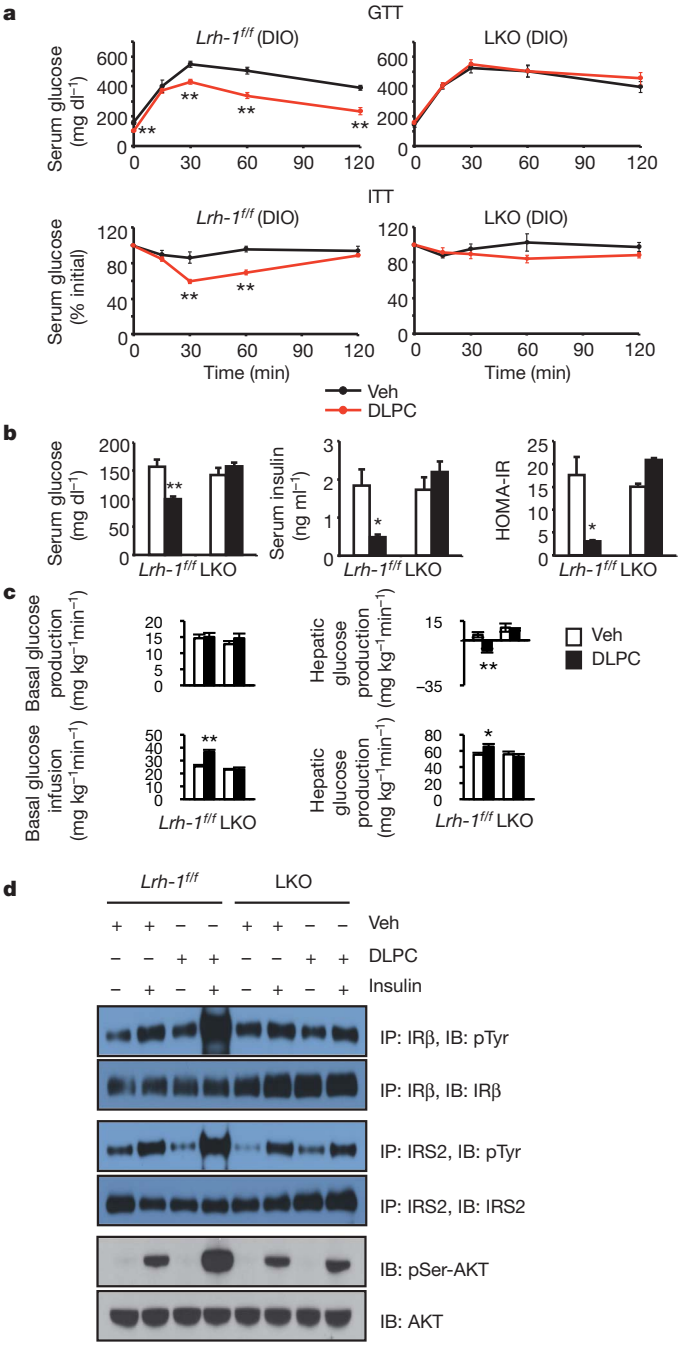


Figure 3 | DLPC improves glucose homeostasis in mouse models of insulin resistance. **a**, Glucose and insulin tolerance were assessed in *Lrh-1*$^{f/f}$ and *Lrh-1*$^{-/-}$ knockout (LKO) DIO mice 2–3 weeks after vehicle or DLPC treatment. **b**, Fasting serum glucose and insulin levels were measured in the same mice shown in **a**. HOMA-IR was calculated from fasting serum glucose and insulin levels. **c**, The high dose (10 mU kg^{-1} min^{-1}) hyperinsulinaemic–euglycaemic clamp (insulin dose of 10 mU kg^{-1} min^{-1}) was used to assess glucose homeostasis in *Lrh-1*$^{f/f}$ DIO mice after 3 weeks of vehicle or DLPC treatment. **d**, Hepatic insulin signalling was examined in *Lrh-1*$^{f/f}$ and LKO DIO mice 2 weeks after vehicle or DLPC treatment. Liver tissue homogenates from 3 mice per group were pooled and immunoprecipitation (IP) and immunoblotting (IB) were as indicated. Results are representative of three independent experiments. Error bars represent mean ± s.e.m. $*P < 0.05$, $**P < 0.01$ versus *Lrh-1*$^{f/f}$ DIO mice treated with vehicle; $n = 4$ mice per group.

glucose homeostasis. These studies indicate that DLPC is a promising therapeutic agent for the treatment of metabolic disorders, and we have initiated a human clinical trial to explore potential beneficial effects in prediabetic patients.

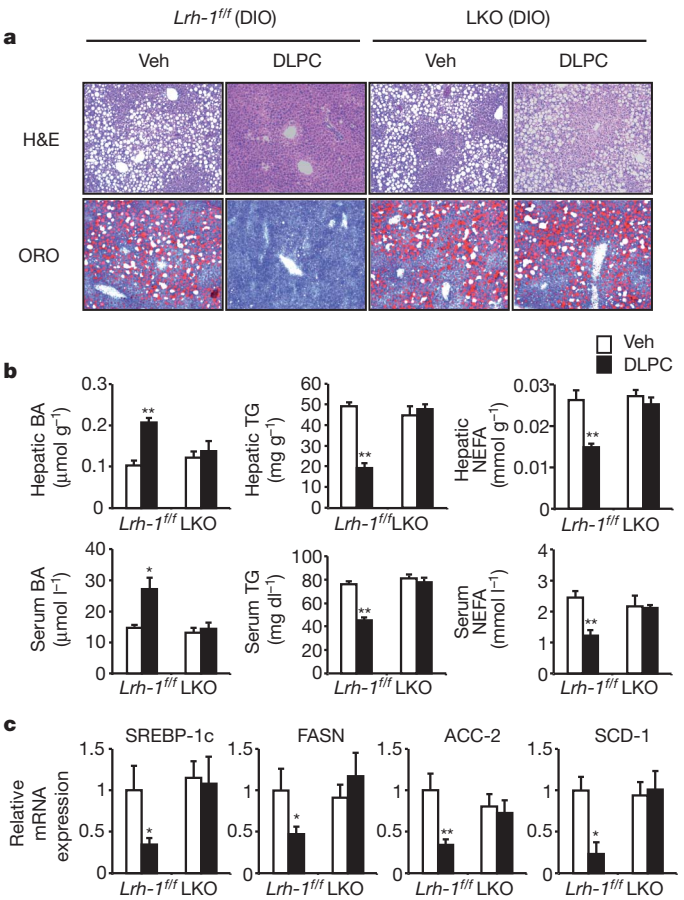


Figure 4 | DLPC reduces liver fat accumulation by suppressing lipogenesis. **a**, Liver sections from *Lrh-1*$^{f/f}$ and LKO DIO mice treated for 3 weeks with vehicle or DLPC were stained with haematoxylin and eosin (H&E) for general morphology or Oil Red O (ORO) for lipid accumulation. Original magnification, ×10. **b**, Hepatic and serum BA, TG and NEFA levels were measured in the same mice described in **a**. **c**, Lipogenic gene expression in the liver was determined using qPCR. mRNA levels are relative to 36B4. Error bars represent mean ± s.e.m. $*P < 0.05$, $**P < 0.01$ versus *Lrh-1*$^{f/f}$ DIO mice treated with vehicle; $n = 4$ mice per group.

METHODS SUMMARY

For transient transfection assays with HeLa, Cos-1 or C3A/HepG2 cells, candidate phospholipids dissolved in ethanol were added to cells for 24 h in medium containing 10% charcoal-treated FBS. Luciferase expression was assayed and normalized using β-galactosidase expression for transfection efficiency. Transfections were done in triplicate. For binding studies, the human LRH-1 ligand-binding domain, residues 291–541, was expressed as a maltose-binding protein fusion protein, cleaved and purified. It was incubated overnight with or without DLPC and specifically bound lipids were extracted with chloroform/methanol and analysed using electrospray mass injection mass spectrometry in the negative-ion mode to detect and identify phospholipids. LanthaScreen binding studies (Invitrogen) used full-length human LRH-1 and PPARγ. For short-term animal studies, C57BL/6 mice were orally gavaged with CA, DPPC, DUPC, or DLPC delivered in a standard vehicle every 12 h for a total of five treatments. Mice were killed 4 h after the final treatment on the morning of day 3. *Lrh-1* liver-specific knockout mice were generated as previously described[2]. For diabetes experiments, *db/db* mice were treated with vehicle or DLPC for 3 weeks. The GTT was performed 2 weeks after treatment. After an additional 1 week treatment, the ITT was performed. Eight-to-ten-week-old male control *Lrh-1*$^{f/f}$ or *Lrh-1*$^{-/-}$ mice were placed on a high-fat diet (45% kcal fat) for 15 weeks. The diet was maintained and mice were treated with vehicle or DLPC by oral gavage. GTT was performed in 18 h fasted mice after 2-week treatments. One week later, ITT was performed in *ad libitum* fed mice. Hyperinsulinaemic clamp (insulin dose of 10 mU kg^{-1} min^{-1}) was performed in the Diabetes and Endocrinology Research Center at the Baylor College of Medicine. All animal experiments were performed according to procedures approved by the Baylor College of Medicine's Institutional Animal Care and Use Committee.

Statistics. Numbers of mice for each group used in experiments are indicated in the figure legends. Statistical analyses were performed with the two-tailed Student's *t*-test, and error bars represent means ± s.e.m. P value < 0.05 was considered statistically significant.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 19 March 2010; accepted 14 March 2011.
Published online 25 May 2011.

1. Mataki, C. *et al.* Compromised intestinal lipid absorption in mice with a liver-specific deficiency of the liver receptor homolog 1. *Mol. Cell. Biol.* **27,** 8330–8339 (2007).
2. Lee, Y. K. *et al.* Liver receptor homolog-1 regulates bile acid homeostasis but is not essential for feedback regulation of bile acid synthesis. *Mol. Endocrinol.* **22,** 1345–1356 (2008).
3. Krylova, I. N. *et al.* Structural analyses reveal phosphatidyl inositols as ligands for the NR5 orphan receptors SF-1 and LRH-1. *Cell* **120,** 343–355 (2005).
4. Ortlund, E. A. *et al.* Modulation of human nuclear receptor LRH-1 activity by phospholipids and SHP. *Nature Struct. Mol. Biol.* **12,** 357–363 (2005).
5. Wang, W. *et al.* The crystal structures of human steroidogenic factor-1 and liver receptor homologue-1. *Proc. Natl Acad. Sci. USA* **102,** 7505–7510 (2005).
6. Cusi, K. Nonalcoholic fatty liver disease in type 2 diabetes mellitus. *Curr. Opin. Endocrinol. Diabetes Obes.* **16,** 141–149 (2009).
7. Watanabe, M. *et al.* Bile acids lower triglyceride levels via a pathway involving FXR, SHP, and SREBP-1c. *J. Clin. Invest.* **113,** 1408–1418 (2004).
8. Lee, Y. K., Parker, K. L., Choi, H. S. & Moore, D. D. Activation of the promoter of the orphan receptor SHP by orphan receptors that bind DNA as monomers. *J. Biol. Chem.* **274,** 20869–20873 (1999).
9. Gu, P. *et al.* Orphan nuclear receptor LRH-1 is required to maintain Oct4 expression at the epiblast stage of embryonic development. *Mol. Cell. Biol.* **25,** 3492–3505 (2005).
10. Lee, Y. K., Choi, Y. H., Chua, S., Park, Y. J. & Moore, D. D. Phosphorylation of the hinge domain of the nuclear hormone receptor LRH-1 stimulates transactivation. *J. Biol. Chem.* **281,** 7850–7855 (2006).
11. Chalkiadaki, A. & Talianidis, I. SUMO-dependent compartmentalization in promyelocytic leukemia protein nuclear bodies prevents the access of LRH-1 to chromatin. *Mol. Cell. Biol.* **25,** 5095–5105 (2005).
12. Chakravarthy, M. F. *et al.* Identification of a physiologically relevant endogenous ligand for PPARα in liver. *Cell* **138,** 476–488 (2009).
13. Li, Y. *et al.* Crystallographic identification and functional characterization of phospholipids as ligands for the orphan nuclear receptor steroidogenic factor-1. *Mol. Cell* **17,** 491–502 (2005).
14. Sablin, E. P. *et al.* Structure of SF-1 bound by different phospholipids: evidence for regulatory ligands. *Mol. Endocrinol.* **23,** 25–34 (2009).
15. Kang, H. W., Wei, J. & Cohen, D. E. PC-TP/StARD2: Of membranes and metabolism. *Trends Endocrinol. Metab.* **21,** 449–456 (2010).
16. Hunt, A. N. Dynamic lipidomics of the nucleus. *J. Cell. Biochem.* **97,** 244–251 (2006).
17. Whitby, R. J. *et al.* Identification of small molecule agonists of the orphan nuclear receptors liver receptor homolog-1 and steroidogenic factor-1. *J. Med. Chem.* **49,** 6652–6655 (2006).
18. Li, D. *et al.* Cyclic AMP-stimulated interaction between steroidogenic factor-1 and diacylglycerol kinase θ facilitates induction of CYP17. *Mol. Cell Biol.* **27,** 6669–6685 (2007).
19. Urs, A. N., Dammer, E. & Sewer, M. B. Sphingosine regulates the transcription of CYP17 by binding to steroidogenic factor-1. *Endocrinology* **147,** 5249–5258 (2006).
20. Lee, H. *et al.* Role for peroxisome proliferator-activated receptor α in oxidized phospholipid-induced synthesis of monocyte chemotactic protein-1 and interleukin-8 by endothelial cells. *Circ. Res.* **87,** 516–521 (2000).
21. McIntyre, T. M. *et al.* Identification of an intracellular receptor for lysophosphatidic acid (LPA): LPA is a transcellular PPARγ agonist. *Proc. Natl Acad. Sci. USA* **100,** 131–136 (2003).
22. Amemiya-Kudo, M. *et al.* Promoter analysis of the mouse sterol regulatory element-binding protein-1c gene. *J. Biol. Chem.* **275,** 31078–31085 (2000).
23. Kanayama, T. *et al.* Interaction between sterol regulatory element-binding proteins and liver receptor homolog-1 reciprocally suppresses their transcriptional activities. *J. Biol. Chem.* **282,** 10290–10298 (2007).
24. Shimomura, I. *et al.* Insulin selectively increases SREBP-1c mRNA in the livers of rats with streptozotocin-induced diabetes. *Proc. Natl Acad. Sci. USA* **96,** 13656–13661 (1999).
25. McGarry, J. D. What if Minkowski had been ageusic? An alternative angle on diabetes. *Science* **258,** 766–770 (1992).
26. Li, S., Brown, M. S. & Goldstein, J. L. Bifurcation of insulin signaling pathway in rat liver: mTORC1 required for stimulation of lipogenesis, but not inhibition of gluconeogenesis. *Proc. Natl Acad. Sci. USA* **107,** 3441–3446 (2010).
27. Kim, K. H. Regulation of mammalian acetyl-coenzyme A carboxylase. *Annu. Rev. Nutr.* **17,** 77–99 (1997).
28. Savage, D. B. *et al.* Reversal of diet-induced hepatic steatosis and hepatic insulin resistance by antisense oligonucleotide inhibitors of acetyl-CoA carboxylases 1 and 2. *J. Clin. Invest.* **116,** 817–824 (2006).
29. Dong, B. *et al.* Activation of nuclear receptor CAR ameliorates diabetes and fatty liver disease. *Proc. Natl Acad. Sci. USA* **106,** 18831–18836 (2009).
30. Dobrzyn, P. *et al.* Stearoyl-CoA desaturase 1 deficiency increases fatty acid oxidation by activating AMP-activated protein kinase in liver. *Proc. Natl Acad. Sci. USA* **101,** 6409–6414 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. A. Kliewer and D. J. Mangelsdorf (UT Southwestern Medical Center) for the gift of *Lrh-1*$^{f/f}$ mice, A. J. Cooney for the OCT4 promoter constructs, C. Mills and D. Kuruvilla for experimental assistance, the Baylor College of Medicine Diabetes Endocrine Research Center (supported by NIH DK-079638 and USDA ARS 6250-52000-055) and the services of the Mouse Metabolism Core for hyperinsulinaemic clamp studies, and the current and previous members of the D.D.M. laboratory for discussions and technical support. Supported by NIH R01 DK068804, the Alkek Foundation and the Robert R. P. Doherty Jr—Welch Chair in Science to D.D.M., and NIH R01 CA134873 to P.R.G.

Author Contributions J.M.L. designed and executed the experiments, interpreted data and co-wrote the manuscript. Y.K.L. and J.L.M. helped with experiments. S.A.B. and P.R.G. performed the fluorescence binding experiments, and M.C.P. and E.A.O. performed the mass spectrometry experiment. D.D.M. supervised the design and interpretation of the experiments and co-wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.D.M. (moore@bcm.edu).

METHODS

Materials. Phospholipids were purchased from Avanti Polar Lipids; fatty acids, CA and CDCA from Sigma-Aldrich; cell culture media and supplements from Invitrogen; insulin from Eli Lilly and Co.; human LRH-1 antibody from R&D systems, antibody against IRβ, IRS2, pSer-AKT and AKT was from Cell Signaling Technology, anti-phosphtyrosine antibody from Millipore; Ad5-CMV–GFP or Cre virus was prepared by the Vector Development Laboratory at the Baylor College of Medicine.

Cell culture and transient transactivation assays. HeLa, Cos-1 and C3A/HepG2 cells were cultured in DMEM containing 10% FBS and 1% penicillin/streptomycin antibiotics. 70–80% confluent cells were replated into a 24-well plate with a 1:5 ratio 24 h before transfection. Cell culture media were changed 1 h before transfection (calcium phosphate method). At 16 h after transfection, candidate phospholipids dissolved in ethanol were added to cells for 24 h in medium containing 10% charcoal treated FBS. Luciferase expression was assayed and normalized using β-galactosidase expression for transfection efficiency. Transfections were done in triplicate. Plasmids used were pcDNA3 for empty vector (100 ng per well), LRH-1/SF-1 luciferase reporter (200 ng per well), actin-β-galactosidase for internal control (150 ng per well), human LRH-1 expression plasmid (100 ng per well), mutant human LRH-1 expression plasmids (100 ng per well, phosphorylation mutant: S238, 243A; sumoylation mutant: K270R; ligand-binding mutant: F342W, I426W), Oct4-PP luciferase reporter (200 ng per well), and Oct4-PP_{mut} luciferase (200 ng per well). Expression plasmids for receptors (100 ng per well) and their cognate luciferase reporters (200 ng per well) used were: human T_3Rβ, TK-28T-Luc; human RXRα, TK-CRBPII-Luc; human RARβ, TK-DR5-Luc; mouse PPARα, mouse PPARδ, mouse PPARγ, TK-PPRE × 3-Luc; mouse FXRα, human FXRα, MMTV-TK-ECRE × 5-Luc; human LXRα, TK-LXRE × 3-Luc; human ERα, TK-ERE-Luc; mouse CAR, human CAR, human PXR, TK-DR4-Luc; mouse SF-1, mouse LRH-1, human LRH-1, LRH-1/SF-1 Luc. For mammalian two-hybrid assays, replated HeLa cells in a 24-well plate were transfected with VP-16 (50 ng per well), VP16-human LRH-1 ligand-binding domain (50 ng per well), Gal4-SRC-3 RID (100 ng per well), G5-TK-Luc (200 ng per well), and actin-β-galactosidase plasmids (150 ng per well). For siRNA experiments, C3A/HepG2 cells were maintained with MEM containing 10% FBS, 1 mM sodium pyruvate, 0.1 mM nonessential amino acids, and 1.5 g l^{-1} sodium bicarbonate. C3A/HepG2 cells were replated into either a 24-well plate for luciferase assay or a 6-well plate for target gene expression and then transfected with human LRH-1 (Dharmacon, ON-TARGETplus SMARTpool, L-003430-00; J-0003430-06: AUAUGAAUAGCCCA UUAUGUU, J-0003430-07: UAGCUGUCCAAAUUUCUCUUU, J-0003430-08: AGGAUUAAGAGCUCACUCCUU, J-0003430-09: UCACCUGAGACAUGGC UUCUU) or control siRNA pool (Dharmacon, si*CONTROL* non-targeting siRNA pool, D-001206-13-05; UAGCGACUAAACACAUCAA, UAAGGCUAU GAAGAGAUAC, AUGUAUUGGCCUGUAUUAG, AUGAACGUGAAUUGC UCAA) using FuGENE 6 (Roche); For the luciferase assay, C3A/HepG2 cells were transfected with LRH-1/SF-1 luciferase and actin-β-galactosidase along with the siRNA pool. Twenty-four hours later, ligands were added, and cells were harvested 24 h later. To check for knockdown, cell extracts were analysed by immunoblot using an antibody against human LRH-1.

***In vitro* binding assays.** GST pull-down was used to examine the interaction between human LRH-1 ligand binding domain and SRC-3 *in vitro* as described[4]. GST alone and GST human LRH-1 ligand-binding domain (amino acid residues, 185–541) were expressed in DH5α strain *E. coli* with 0.5 mM IPTG for 4 h and then purified with glutathione-sepharose beads (GE Healthcare). GST proteins were incubated overnight at 4 °C in 50 mM Tris-HCL (pH 7.6), 0.2% Tween-20, 100 mg ml^{-1} BSA and 300 mM NaCl with various phospholipids. Full-length [^{35}S]methionine-labelled SRC-3 proteins (2 μl) were added to each reaction and incubated for 2 h at 4 °C. Unbound and nonspecific proteins were removed by washing five times with the same buffer. Specifically bound proteins were eluted by treatment with SDS sample buffer, subjected to SDS–PAGE, and visualized by autoradiography. The amount of specifically bound SRC-3 proteins was determined by densitometry (Personal Densitometer SI; Molecular Dynamics).

Lanthascreen assays were as described by the manufacturer (Invitrogen) using full-length human LRH-1 and a fluorescein-tagged SRC-2 coactivator peptide, and PPARγ and Fluoromone[31].

For mass spectrometry, the human LRH-1 ligand-binding domain (LBD), residues 291–541, was expressed as a maltose-binding protein fusion protein, cleaved and purified as described previously[4]. The pure protein was stored in a final buffer containing 150 mM NaCl, 20 mM HEPES and 5% glycerol. For binding studies, DLPC or DPPC dissolved in ethanol was evaporated in a clean glass cuvette at 50 °C under a stream nitrogen gas. Two millilitres of buffer containing 150 mM NaCl, 20 mM HEPES (pH = 7.5) and 5% glycerol was added to the cuvette containing dried DLPC or DPPC and was sonicated until the solution was optically clear. Human LRH-1 LBD was then added to the DLPC or DPPC vesicles at a ratio of 1:1 or 1:5 (human LRH-1 LBD:PC). The mixture was incubated for one hour at 37 °C followed by 24 h at 11 °C. The human-LRH-1–lipid complex was then purified by size exclusion chromatography to remove unbound phospholipids. Protein purity was assed by SDS–PAGE. Bound lipids were analysed using electrospray mass injection mass spectrometry (ESI-MS) in the negative-ion mode to detect and identify phospholipids. Approximately 6 mg of human LRH-1 LBD or the human LRH-1-LBD–lipid complexes were extracted with a 2:1 chloroform/methanol solution, diluted in 200 ml chloromethylene and analysed by negative-ion ESI-MS on a Thermo LTQ FTMS using direct injection analysis with electrospray ionization. The high-resolution analyses were performed in the FTMS at a resolution of 100,000 at 400 *m/z*. The MS/MS experiments were done in the ion trap portion of the instrument with a mass selection of 3 AMU and a normalized collision energy of 30 V. The major phospholipid species were identified by accurate mass measurements and MS/MS via collisionally induced dissociation (CID), which yields product ions characteristic of the head groups and attached fatty acids.

Animal studies. C57BL/6 mice and *db/db* mice were purchased from Harlan laboratories. Eight-week-old male C57BL/6 mice were orally gavaged with CA, DPPC, DUPC, or DLPC at a dose of 100 mg kg^{-1} body weight, delivered in a standard vehicle for delivery of hydrophobic compounds (4:1 of PEG-400 and Tween-80) every 12 h for a total of five treatments. Mice were killed 4 h after the final treatment on the morning of day 3. Harvested tissues were immediately frozen in liquid nitrogen for molecular studies. Twelve-week-old male *db/db* mice were used for diabetes studies. *db/db* mice were given compounds (vehicle or DLPC, $n = 5$ mice per group) at the dose of 100 mg kg^{-1} day^{-1}. After 2 weeks of treatments GTT (1.5 g kg^{-1} intraperitoneal injection) was performed in 18 h fasted mice. Treatments were continued for an additional week, and ITT (2 U kg^{-1} intraperitoneal injection) was performed in *ad libitum* fed mice. Serum insulin levels were determined using Rat/Mouse Insulin ELISA Kit from Linco Research. *Lrh-1$^{f/f}$* mice[2] were maintained on mixed C57BL6/129 backgrounds and were given by the Kliewer/Mangelsdorf laboratory. In brief, 4-month-old male *Lrh-1$^{f/f}$* littermates were tail vein injected with either Ad5-CMV–GFP (3×10^9 p.f.u.) or Ad5-CMV-Cre (3×10^9 p.f.u.). For acute experiments, these mice were orally gavaged daily with each compound starting 2 weeks after adenovirus injection as described in C57BL/6 mice above. Liver-specific *Lrh-1* ablation (LKO) was also achieved by crossing *Lrh-1$^{f/f}$* mice with albumin-Cre transgenic mice obtained from the O'Malley laboratory at the Baylor College of Medicine. To confirm tissue-specific deletion of exon 5 of *Lrh-1*, genomic DNA was extracted from tail, liver and intestine, and PCR analysis was performed as shown previously[2]. For diabetes experiments, 8–10-week-old male control *Lrh-1$^{f/f}$* or LKO mice were placed on a high-fat diet (Research diets; 45% kcal fat) for 15 weeks. The diet was maintained and mice were treated with vehicle or DLPC (dose of 100 mg kg^{-1} day^{-1}) by oral gavage. GTT (2 g kg^{-1} intraperitoneal injection) was performed in 18 h fasted mice after 2-week treatments. 1 week later, ITT (1 U kg^{-1}, intraperitoneal injection) was performed in *ad libitum* fed mice. Glucose levels were analysed using a glucometer (LifeScan). Insulin resistance (HOMA-IR) was calculated as following: fasting glucose (mg dl^{-1}) × fasting insulin (μU ml^{-1}) / 405. Hyperinsulinaemic clamp (insulin dose of 10 mU kg^{-1} min^{-1}) was performed and calculated as described in our previous publication[32]. Ad–GFP- or Ad-Cre-infected *Lrh-1$^{f/f}$* mice fed the high-fat diet for 15 weeks were used for the diabetes study as shown above. Mice were housed in a temperature-controlled room in pathogen-free facilities on a 12 h light/dark cycle (07:00 on, 19:00 off) and had free access to water and standard chow diet. All animals received humane care according to the criteria outlined in the "Guide for the Care and Use of Laboratory Animals" prepared by the National Academy of Sciences and published by the National Institutes of Health.

RNA isolation and mRNA quantification. Total RNA was isolated from C3A/HepG2 cells or snap-frozen liver tissues using Trizol Reagent (Invitrogen) and prepared for the cDNA with QuantiTect reverse transcriptase (Qiagen). Hepatic gene expression ($n = 4$–5) was determined by qPCR using FastStart SYBR Green master (ROX) (Roche). mRNA levels were normalized by the 36B4 gene. Primer information can be provided upon request.

Serum and tissue lipid analysis. Blood was collected from the orbital plexus and transferred into gel/clot activator tubes (Terumo). Samples were centrifuged at 6,000*g* for 5 min to separate serum. To extract bile acids from liver or intestine, each tissue was weighed and homogenized in 75% ethanol. The homogenate was incubated at 50 °C for 2 h to extract bile acids and centrifuged at 6,000*g* for 10 min at 4 °C. The bile acid content of the supernatant was determined and normalized with tissue weight used. To extract other lipids, snap-frozen liver fragments were weighed and homogenized in nine volumes of PBS. Two-hundred microlitres of the homogenate was transferred into 1,200 μl of chloroform:methanol (2:1; v/v) mixture and mixed vigorously for 30 s. One-hundred microlitres of PBS was then added, and the resulting suspension was mixed vigorously for 15 s then centrifuged at 4,200*g* for 10 min at 4 °C. Two-hundred microlitres of the chloroform:methanol layer (bottom phase) was transferred into a tube and evaporated for dryness. The

dried lipid residue was resuspended in 100 µl of 1% Triton X100 in absolute ethanol for 4 h with constant rotation. Bile acids levels were measured using the bile acid L3K assay kit (Diagnostic Chemicals). Cholesterol and triglyceride levels were determined by assay kits from Thermo DMA. Free fatty acids were assayed using a kit obtained from WAKO Chemicals.

***In vivo* insulin stimulation and analysis of insulin signalling.** Mice were fasted overnight and injected intraperitoneally with insulin (1 U kg^{-1}) or PBS. Five minutes after injection, tissues were removed, frozen in liquid nitrogen, and stored at −80 °C until use. For protein extraction, tissues were homogenized in a cold lysis buffer (50 mM Tris-HCl, pH 7.4; 1% NP-40; 0.5% sodium deoxycholate; 150 mM NaCl; 1 mM EDTA) containing protease and phosphatase inhibitor cocktails (Roche). After homogenization, the tissue lysates were allowed to solubilize for 1 h at 4 °C with rotation, and then were centrifuged at 19,700*g* for 30 min at 4 °C. The supernatants were used for immunoprecipitations and immunoblot analyses of insulin signalling proteins.

Histology. Liver was removed and pieces were fixed in 10% (v/v) neutralized formalin solution (J. T. Baker), embedded in paraffin, sectioned at 5 µm, and stained with haematoxylin and eosin. For Oil Red O staining, frozen liver tissues embedded in O.C.T. compound (Tissue-Tek) were used. Histological analysis performed in the Comparative Pathology Laboratory at Baylor College of Medicine.

Statistics. Numbers of mice for each group used in experiments are indicated in the figure legends. Statistical analyses were performed with the two-tailed Student's *t*-test, and error bars represent means ± s.e.m. P value < 0.05 was considered statistically significant.

31. Vidović, D., Busby, S. A., Griffin, P. R. & Schurer, S. C. A combined ligand- and structure-based virtual screening protocol identifies submicromolar PPARγ partial agonists. *ChemMedChem* **6,** 94–103 (2011).
32. Ma, K., Saha, P. K., Chan, L. & Moore, D. D. Farnesoid X receptor is essential for normal glucose homeostasis. *J. Clin. Invest.* **116,** 1102–1109 (2006).

LETTER

doi:10.1038/nature10117

SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx

Nadine Laguette[1], Bijan Sobhian[1], Nicoletta Casartelli[2], Mathieu Ringeard[1], Christine Chable-Bessia[1], Emmanuel Ségéral[3], Ahmad Yatim[1], Stéphane Emiliani[3], Olivier Schwartz[2] & Monsef Benkirane[1]

The primate lentivirus auxiliary protein Vpx counteracts an unknown restriction factor that renders human dendritic and myeloid cells largely refractory to HIV-1 infection[1–6]. Here we identify SAMHD1 as this restriction factor. SAMHD1 is a protein involved in Aicardi–Goutières syndrome, a genetic encephalopathy with symptoms mimicking congenital viral infection, that has been proposed to act as a negative regulator of the interferon response[7]. We show that Vpx induces proteasomal degradation of SAMHD1. Silencing of *SAMHD1* in non-permissive cell lines alleviates HIV-1 restriction and is associated with a significant accumulation of viral DNA in infected cells. Concurrently, overexpression of SAMHD1 in sensitive cells inhibits HIV-1 infection. The putative phosphohydrolase activity of SAMHD1 is probably required for HIV-1 restriction. Vpx-mediated relief of restriction is abolished in SAMHD1-negative cells. Finally, silencing of *SAMHD1* markedly increases the susceptibility of monocytic-derived dendritic cells to infection. Our results demonstrate that SAMHD1 is an antiretroviral protein expressed in cells of the myeloid lineage that inhibits an early step of the viral life cycle.

Most monocytic cell lines fail to recapitulate the HIV-1 restriction phenotype that is witnessed in primary dendritic cells and to a lesser extent in macrophages. One exception is represented by phorbol 12-myristate 13-acetate (PMA)-differentiated THP-1 monocytic cells[1,2,4–6]. Transduction of differentiated THP-1 cells by virus-like particles containing Vpx (VLP-Vpx) has been shown to increase their permissiveness to HIV-1 infection[1]. To identify the restriction factor targeted by Vpx, we generated a stable THP-1 cell line expressing Flag-HA epitope-tagged Vpx_{mac251} from sooty mangabey (THP-1-Vpx). As expected, THP-1-Vpx cells were 17-fold more permissive to HIV-1 infection than parental cells when exposed to a vesicular stomatitis virus G protein (G)-pseudotyped HIV-1 with the luciferase gene in place of *nef* (HIV-LUC-G) (Supplementary Fig. 1). After differentiation, extracts were prepared from THP-1 and THP-1-Vpx cells. Flag-HA–Vpx_{mac251} (F/H-Vpx) was purified using tandem affinity chromatography[8]. Purified Vpx_{mac251}-associated proteins were resolved by SDS–polyacrylamide gel electrophoresis (PAGE) and silver stained (Fig. 1a). The eluates were further analysed by mass spectrometry to allow for the identification of cellular partners that engage with Vpx_{mac251} in non-permissive cells. Previously described Vpx_{mac251} interactants were recovered, including the DDB1–CUL4–DCAF E3 ligase complex[9,10], confirming the validity of our approach. The SAM- and HD-domain-containing protein SAMHD1 was identified as a major Vpx_{mac251}-interacting protein (Table 1). SAMHD1 was initially isolated as an interferon (IFN)-γ-induced factor in macrophages and dendritic cells[11,12] for which a function in innate immunity has been suggested[7,13]. Furthermore, mutations in *SAMHD1* have been shown to be responsible for 5% of Aicardi–Goutières syndrome cases, a genetically heterogeneous disorder characterized by inappropriate activation of the immune system and aberrant IFN-α secretion, symptoms reminiscent of congenital infection[7].

The interaction of SAMHD1 with Vpx_{mac251} was confirmed by Flag-immunoprecipitation of F/H-Vpx and western blot analysis using a SAMHD1-specific antibody (Fig. 1b). Notably, analysis of

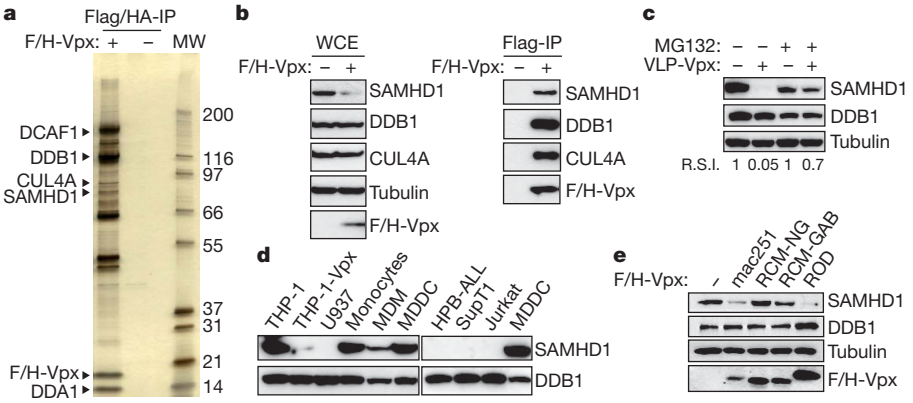


Figure 1 | SAMHD1 interacts with Vpx and is degraded by the proteasome. **a**, SIV_{mac251} Vpx was tandem-affinity-purified from Flag- and HA-tagged Vpx_{mac251} (F/H-Vpx)-expressing THP-1 cells (THP-1-Vpx) and peptide-eluted under native conditions. Eluates were separated on SDS–PAGE and silver stained. SAMHD1 and major, previously described, Vpx_{mac251} interactants identified using tandem mass spectrometry are indicated (MW, protein molecular weight marker in kDa). **b**, Whole-cell extract (WCE) and Flag-immunoprecipitated F/H-Vpx analysis by western blot against DDB1, CUL4A and SAMHD1. **c**, THP-1 cells were treated with 50 μM MG132 for 2 h before a 2-h incubation with Vpx_{mac251} containing virus-like particles (VLP-Vpx). After a further overnight incubation with MG132, whole-cell extracts were prepared and analysed by western blot with the indicated antibodies. R.S.I., relative signal intensity. **d**, Analysis of the expression profile of SAMHD1 in different cell types by western blot. **e**, THP-1 cells were transduced with a bicistronic retroviral vector allowing expression of Vpx_{mac251}, Vpx_{ROD}, $Vpx_{RCM\text{-}NG}$ and $Vpx_{RCM\text{-}GAB}$ and a selectable marker. After cell sorting and whole-cell extraction, the ability of Vpx variants to degrade SAMHD1 in THP-1 whole-cell extract was analysed by western blot using the indicated antibodies.

[1]Institut de Génétique Humaine, Laboratoire de Virologie Moléculaire, CNRS UPR1142, Montpellier 34000 France. [2]Institut Pasteur, Virus and Immunity Unit, URA CNRS 3015, Paris, France. [3]Institut Cochin, Université Paris Descartes, CNRS UMR 8104 INSERM U567, Paris, France.

Table 1 | Major F/H-Vpx interactants identified by mass spectrometry

Protein symbol	Peptide count
DDB1	620
DCAF1	499
SAMHD1	290
CUL4A	55
CUL4B	49
DDA1	18

whole-cell extracts of THP-1-Vpx as compared to THP-1 reveals that expression of Vpx_{mac251} correlates with lower expression levels of SAMHD1 (Fig. 1b). Transient delivery of Vpx_{mac251} into THP-1 cells through VLP-Vpx exposure caused a marked decrease in SAMHD1 levels in THP-1 cells (Fig. 1c). Moreover, treatment of cells with the proteasome inhibitor MG132 restored SAMHD1 protein levels (Fig. 1c), strongly suggesting that Vpx_{mac251} induces proteasomal degradation of SAMHD1. Additionally, when SAMHD1 was expressed in HeLa cells, Vpx also caused its degradation, demonstrating that Vpx-induced degradation of SAMHD1 is not a cell-type-specific process and excluding a potential transcriptional effect of Vpx on *SAMHD1* (Supplementary Fig. 2). These results led us to postulate that SAMHD1 may be the restriction factor that renders dendritic cells refractory to HIV-1 infection. Consistently, SAMHD1 is highly expressed in HIV-1 non-permissive cells such as THP-1, monocytes and monocyte-derived dendritic cells (MDDCs), whereas it is absent from HIV-1-sensitive T-cell lines such as Jurkat, SupT1, human peripheral blood acute lymphoid leukaemia (HPB-ALL) and U937 (Fig. 1d). In correlation with their degree of permissiveness to HIV-1 (refs 1, 4), monocyte-derived macrophages (MDMs) express low levels of SAMHD1 as compared to their highly refractory monocyte precursor (Fig. 1d).

It has been reported that Vpx-mediated enhancement of HIV-1 infection in dendritic cells and myeloid cells is conserved exclusively within the SIV_{SM} (sooty mangabey) and HIV-2 lineage (refs 1, 2 and Supplementary Fig. 2a). Concurrent with this observation, we show that Vpx_{mac251} and HIV-2_{ROD} Vpx (Vpx_{ROD}) both caused degradation of SAMHD1, whereas Vpx from SIV_{RCM} of red-capped mangabeys (isolates of Nigerian (Vpx_{RCM-NG}) or Gabonese ($Vpx_{RCM-GAB}$) origin) failed to degrade SAMHD1 when expressed in THP-1 (Fig. 1e and Supplementary Fig. 3a, b) or HeLa cells (Supplementary Fig. 3c). Additionally, loss-of-function mutants Vpx(Q76A) and Vpx(F80A) also failed to degrade SAMHD1in differentiated THP-1 cells (Supplementary Fig. 3d).

To assess directly the impact of expression of SAMHD1 on HIV-1 infection, we used short hairpin RNAs (shRNA) to generate SAMHD1-silent THP-1 cells (THP-1-shSAMHD1) or scrambled shRNA (THP-1-scr) (Supplementary Fig. 4). Infection of THP-1-shSAMHD1 cells with HIV-LUC-G results in up to a 12-fold increase in luciferase activity as compared to THP-1-scr cells, demonstrating that depletion of SAMHD1 is sufficient to increase the permissiveness of THP-1 cells to HIV-1 infection (Fig. 2a). As expected, HIV-1 infection of THP-1-shSAMHD1 cells was not further enhanced by treatment with VLP-Vpx (Fig. 2b and Supplementary Fig. 5). Importantly, expression of a shRNA-resistant SAMHD1 mutant (SAMHD1-R) in THP-1-shSAMHD1 cells restored the restriction phenotype in these cells (Fig. 2d and Supplementary Fig. 6). To investigate further whether SAMHD1 possesses an intrinsic restriction activity targeting HIV-1, we stably expressed Flag-HA-tagged SAMHD1 (SAMHD1-F/H) in permissive U937 monocytic cells (U937-SAMHD1). Differentiated U937-SAMHD1 cells were 16-fold less permissive to infection with HIV-LUC-G than parental cells (Fig. 2e and Supplementary Fig. 8). HD domains have an important role in nucleotide metabolism through their nucleotidase and phosphodiesterase activities, where H and D residues are of crucial importance[14]. Interestingly, U937 cells stably expressing the SAMHD1(HD/AA) (U937-SAMHD1(HD/AA)) mutant did not show restriction activity towards HIV-1 as compared to U937-SAMHD1 cells, indicating that the putative phosphohydrolase activity of SAMHD1 may be required for restriction of HIV-1

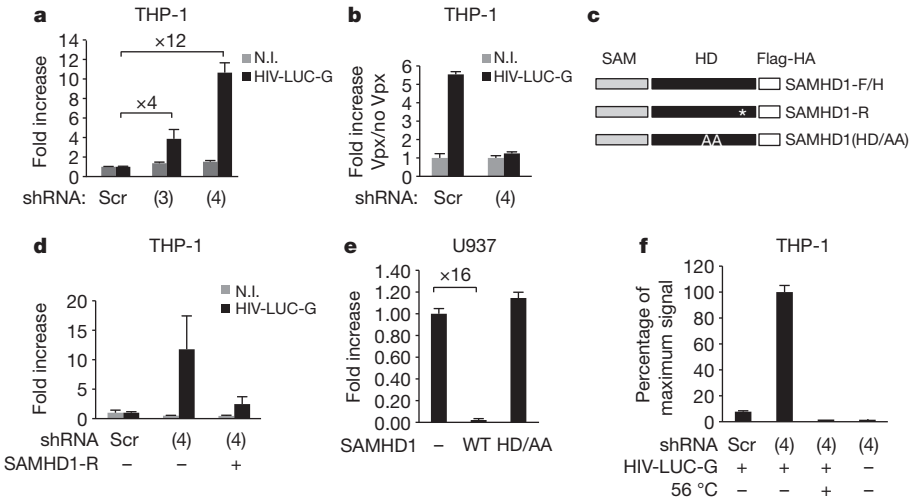


Figure 2 | SAMHD1 restricts HIV-1 infection in THP-1 cells. THP-1 cells were engineered to stably express shRNA 3 (3) or shRNA 4 (4) specifically targeting SAMHD1 (THP-1-shSAMHD1) or scrambled shRNA (THP-1-scr). **a**, THP-1-shSAMHD1 and THP-1-scr cells were infected with 50 ng of HIV-LUC-G. Luciferase activity was measured 24 h after infection and normalized for protein concentration in analysed samples. Results are expressed as fold increase of luciferase activity in THP-1-shSAMHD1 over THP-1-scr cells. N.I., non-infected. **b**, THP-1-shSAMHD1 and THP-1-scr cells were treated with VLP-Vpx before infection with 50 ng of HIV-LUC-G. Luciferase activity was measured as in **a**. Results are expressed as fold increase of luciferase activity in VLP-Vpx-treated over untreated cells. **c**, Mutants of Flag- and HA-tagged SAMHD1 (SAMHD1-F/H) were generated that are either shSAMHD1-resistant (SAMHD1-R) or mutated in the HD domain (SAMHD1(HD/AA)). These mutants were introduced in an MLV expression vector. Asterisk indicates synonymous mutation. **d**, THP-1-shSAMHD1 cells were transduced with SAMHD1-R for 48 h or left untreated, differentiated and infected with 100 ng of HIV-LUC-G. Luciferase activity was measured and expressed as in **a**. **e**, U937 myeloid cells were transduced with SAMHD1-F/H or SAMHD1(HD/AA) for 24 h. After a further 16-h differentiation step, cells were infected with 10 ng of HIV-LUC-G. Luciferase activity was measured as in **a**. Results are expressed as fold increase luciferase activity in transduced over parental U937 cells. **f**, Total viral DNA was quantified by quantitative PCR in THP-1-shSAMHD1 and THP-1-scr cells 24 h after infection with HIV-LUC-G or heat-inactivated virus (56 °C). Results are expressed as per cent maximum signal intensity. All graphs show mean ± standard deviation from a representative experiment ($n = 5$).

(Fig. 2e and Supplementary Figs 7 and 8). Additionally, transient expression of SAMHD1-F/H, but not SAMHD1(HD/AA), in permissive HeLa cells induced restriction of HIV-1 infection (Supplementary Fig. 7).

Vpx has been shown to facilitate HIV-1 replication by promoting accumulation of viral DNA[1,2,4,5,15]. To investigate at which step of the viral replication cycle SAMHD1-dependent restriction operates, we quantified total viral DNA species 24 h after infection of THP-1-shSAMHD1 cells (Fig. 2f). A 13-fold accumulation of total viral DNA was observed in SAMHD1-silenced cells as compared to their THP-1-scr counterpart. This observation locates the restriction operated by SAMHD1 at the reverse transcription step, which has been previously described to be overcome by Vpx in dendritic and myeloid cells[2].

Exposure of differentiated THP-1, MDDCs and MDMs to VLP-Vpx relieves restriction to HIV-1 infection (Fig. 3a and Supplementary Fig. 9) and correlates with a decrease in SAMHD1 levels (Fig. 3b and Supplementary Fig. 9). Lower basal levels of SAMHD1 in primary MDMs (Fig. 1d) may be accountable for the weaker impact of VLP-Vpx treatment of these cells (Supplementary Fig. 9). Next, we asked whether SAMHD1 restricts HIV-1 infection of primary human dendritic cells. Immature MDDCs were prepared from the blood of four healthy donors. To silence SAMHD1 expression, we treated MDDCs with two different SAMHD1-specific siRNAs (si-SAMHD1-1 or si-SAMHD1-2). Scrambled or siRNA targeting dynamin 2 (siDYN2) were used as controls (Fig. 3c and Supplementary Fig. 10). MDDCs were transduced (at 48 h after silencing) with green-fluorescent-protein- or luciferase-encoding lentiviral vectors (LV-GFP or LV-LUC, respectively) or with a VSV-G-pseudotyped HIV (HIV-G). Silencing of SAMHD1 resulted in up to a 6-fold increase of GFP-positive cells with LV-GFP, 25-fold enhanced luciferase activity with LV-LUC, and up to 34-fold increased Gag-positive cells with HIV-G (Fig. 3d–f and Supplementary Fig. 10), demonstrating that SAMHD1 silencing in dendritic cells enhances their susceptibility to HIV-1 infection.

So far, three restriction factors (TRIM5α, APOBEC-3G and tetherin) have been identified that could have constituted a major hindrance to HIV-1 replication, had the virus not developed means to counteract or to escape their action[16–18]. These restriction factors are part of the intrinsic immunity that is circumvented by HIV-1 mainly through the action of its auxiliary proteins. However, the cell-type-specific restriction factor SAMHD1 is not counteracted by HIV-1, resulting in poorly efficient replication in dendritic cells. Indeed, in the normal course of HIV-1 infection and because dendritic cells are non-permissive, these cells rather facilitate viral dissemination through trans-enhancement of infection, eventually favouring $CD4^+$ T-cell depletion[19,20]. Poor HIV-1 replication in dendritic cells may also allow for avoidance of a recently described cryptic viral sensor that would otherwise elicit antiviral interferon-induced immune responses[20,21]. Similarly, in productively infected $CD4^+$ T cells, through the action of the cellular DNase TREX1, HIV-1 avoids the induction of type 1 IFN production that could result from accumulation of viral DNA[20,22]. Of note, both TREX1 and SAMHD1 deficiencies lead to Aicardi–Goutières syndrome, indicating that they have an impact on the same pathway of cell-intrinsic antiviral response[7]. SAMHD1, through the putative nucleotidase activity, may degrade or prevent accumulation of HIV DNA. It will be worth determining further the relative roles of TREX1 and SAMHD1 in the control of virus infection and in the triggering of innate antiviral and inflammatory responses.

Our findings position SAMHD1 as pivotal to the fate of infection by HIV-1 in cells of the myeloid lineage. Thus, modulating SAMHD1 function could render human hosts more prone to develop appropriate innate and adaptive immune responses[19,22–24]. Our findings should be integrated in the development of dendritic-cell-targeted vaccines against HIV/AIDS.

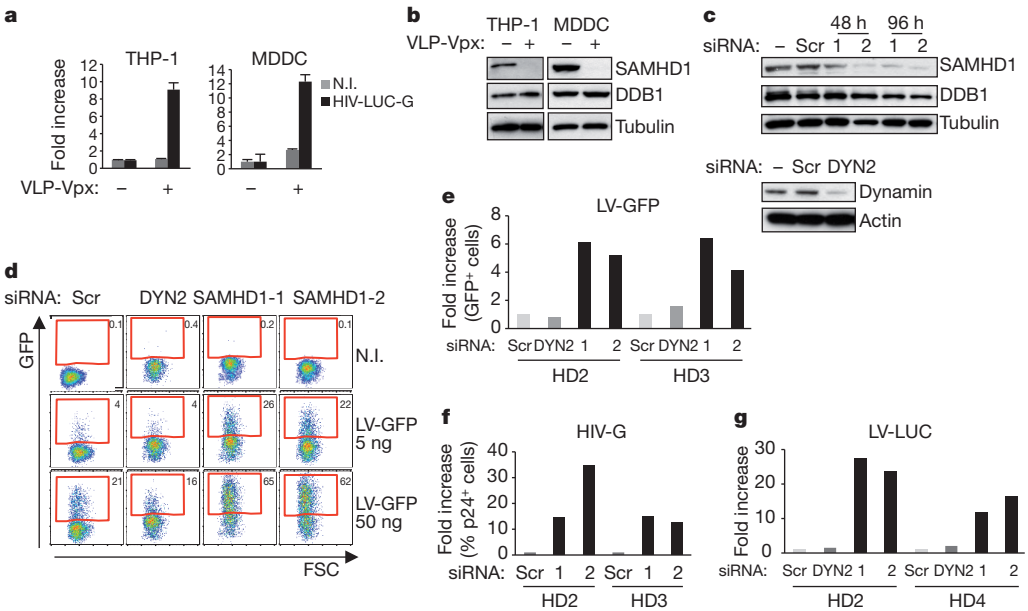


Figure 3 | SAMHD1 restricts HIV-1 infection in primary MDDCs. **a**, THP-1 and monocyte-derived dendritic cells (MDDCs) from healthy donors (HD) were treated with VLP-Vpx for 2 h and infected with 100 ng of HIV-LUC-G. Luciferase activity was measured at 48 h after infection and normalized for protein concentration. Results are expressed as fold increase luciferase activity in VLP-Vpx treated over untreated cells. Graphs show mean ± standard deviation from a representative experiment ($n = 4$). **b**, Cells from **a** were subjected to whole-cell extraction before analysis by western blot using indicated antibodies. **c**, MDDCs from HD1 were mock transfected or transfected with siRNA targeting *SAMHD1* (siRNA 1 and 2) or dynamin 2 (DYN2) for 48 and 96 h before whole-cell extraction and analysis by western blot using indicated antibodies. **d**, Cells from HD2 were transduced at 48 h with GFP encoding lentiviral vector (LV-GFP, 50 ng) and analysed by flow cytometry after 4 days. **e**, HD2 and HD3 were transduced as in **d**, except that a concentration of 5 ng of LV-GFP was used. The percentage of GFP-positive cells from HD2 and HD3 was quantified and expressed as fold increase GFP-positive cells relative to scrambled siRNA-treated cells. **f**, Transfected MDDCs from HD2 and HD3 were infected with 10 ng of HIV-G. Results are expressed as fold increase of p24-positive cells relative to scrambled siRNA-treated cells. **g**, MDDCs from HD2 and HD4 were transfected as in **c** before infection with 10 ng of luciferase-expressing lentiviral vector (LV-LUC). Results are expressed as in Fig. 2a.

METHODS SUMMARY

Vpx_{mac251} was purified from 5×10^9 differentiated THP-1 cells stably expressing Flag- and HA-tagged Vpx_{mac251} (F/H-Vpx) by two-step affinity chromatography according to the standard method[8]. Extracts were first incubated with anti-Flag antibody conjugated agarose beads and the bound polypeptides were eluted with Flag peptide under native conditions. The Flag affinity-purified material was further immunopurified by affinity chromatography using anti-HA antibody-conjugated agarose beads and eluted under native conditions using HA peptide. Five per cent of immunoaffinity-purified F/H-Vpx or mock were resolved on SDS–PAGE and silver stained while the remainder was stained with Coomassie-R250. Regions of the gel were excised and subsequently analysed by tandem mass spectrometry. U937 and THP-1 cells were differentiated overnight with 30 ng ml^{-1} PMA before all infections. MDDCs and MDMs were prepared from human buffy coats of healthy donors. Silencing of SAMHD1 in THP-1 and in MDDCs[25] was achieved by selecting cells stably expressing shRNA or using siRNA targeting *SAMHD1*. All experiments were repeated at least four times.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 10 March 2010; accepted 18 April 2011.
Published online 25 May 2011.

1. Goujon, C. *et al.* Characterization of simian immunodeficiency virus SIVSM/human immunodeficiency virus type 2 Vpx function in human myeloid cells. *J. Virol.* **82,** 12335–12345 (2008).
2. Goujon, C. *et al.* SIV_{SM}/HIV-2 Vpx proteins promote retroviral escape from a proteasome-dependent restriction pathway present in human dendritic cells. *Retrovirology* **4,** 2 (2007).
3. Hirsch, V. M. *et al.* Vpx is required for dissemination and pathogenesis of SIV_{SM} PBj: evidence of macrophage-dependent viral amplification. *Nature Med.* **4,** 1401–1408 (1998).
4. Kaushik, R., Zhu, X., Stranska, R., Wu, Y. & Stevenson, M. A cellular restriction dictates the permissivity of nondividing monocytes/macrophages to lentivirus and gammaretrovirus infection. *Cell Host Microbe* **6,** 68–80 (2009).
5. Sharova, N. *et al.* Primate lentiviral Vpx commandeers DDB1 to counteract a macrophage restriction. *PLoS Pathog.* **4,** e1000057 (2008).
6. Ayinde, D., Maudet, C., Transy, C. & Margottin-Goguet, F. Limelight on two HIV/SIV accessory proteins in macrophage infection: is Vpx overshadowing Vpr? *Retrovirology* **7,** 35 (2010).
7. Rice, G. I. *et al.* Mutations involved in Aicardi-Goutieres syndrome implicate SAMHD1 as regulator of the innate immune response. *Nature Genet.* **41,** 829–832 (2009).
8. Nakatani, Y. & Ogryzko, V. Immunoaffinity purification of mammalian protein complexes. *Methods Enzymol.* **370,** 430–444 (2003).
9. Bergamaschi, A. *et al.* The human immunodeficiency virus type 2 Vpx protein usurps the CUL4A–DDB1 DCAF1 ubiquitin ligase to overcome a postentry block in macrophage infection. *J. Virol.* **83,** 4854–4860 (2009).
10. Srivastava, S. *et al.* Lentiviral Vpx accessory factor targets VprBP/DCAF1 substrate adaptor for cullin 4 E3 ubiquitin ligase to enable macrophage infection. *PLoS Pathog.* **4,** e1000059 (2008).
11. Li, N., Zhang, W. & Cao, X. Identification of human homologue of mouse IFN-γ induced protein from human dendritic cells. *Immunol. Lett.* **74,** 221–224 (2000).
12. Liao, W., Bao, Z., Cheng, C., Mok, Y. K. & Wong, W. S. Dendritic cell-derived interferon-γ-induced protein mediates tumor necrosis factor-α stimulation of human lung fibroblasts. *Proteomics* **8,** 2640–2650 (2008).
13. Zhao, D., Peng, D., Li, L., Zhang, Q. & Zhang, C. Inhibition of G1P3 expression found in the differential display study on respiratory syncytial virus infection. *Virol. J.* **5,** 114 (2008).
14. Zimmerman, M. D., Proudfoot, M., Yakunin, A. & Minor, W. Structural insight into the mechanism of substrate specificity and catalytic activity of an HD-domain phosphohydrolase: the 5′-deoxyribonucleotidase YfbR from *Escherichia coli*. *J. Mol. Biol.* **378,** 215–226 (2008).
15. Gramberg, T., Sunseri, N. & Landau, N. R. Evidence for an activation domain at the amino terminus of simian immunodeficiency virus Vpx. *J. Virol.* **84,** 1387–1396 (2010).
16. Stremlau, M. *et al.* The cytoplasmic body component TRIM5α restricts HIV-1 infection in Old World monkeys. *Nature* **427,** 848–853 (2004).
17. Sheehy, A. M., Gaddis, N. C., Choi, J. D. & Malim, M. H. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* **418,** 646–650 (2002).
18. Neil, S. J., Zang, T. & Bieniasz, P. D. Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. *Nature* **451,** 425–430 (2008).
19. Altfeld, M., Fadda, L., Frleta, D. & Bhardwaj, N. DCs and NK cells: critical effectors in the immune response to HIV-1. *Nature Rev. Immunol.* **11,** 176–186 (2011).
20. Yan, N. & Lieberman, J. Gaining a foothold: how HIV avoids innate immune recognition. *Curr. Opin. Immunol.* **23,** 21–28 (2011).
21. Manel, N. *et al.* A cryptic sensor for HIV-1 activates antiviral innate immunity in dendritic cells. *Nature* **467,** 214–217 (2010).
22. Borrow, P., Shattock, R. J. & Vyakarnam, A. Innate immunity against HIV: a priority target for HIV prevention research. *Retrovirology* **7,** 84 (2010).
23. Cobb, A. *et al.* Development of a HIV-1 lipopeptide antigen pulsed therapeutic dendritic cell vaccine. *J. Immunol. Methods* **365,** 27–37 (2011).
24. Lepelley, A. *et al.* Innate sensing of HIV-infected cells. *PLoS Pathog.* **7,** e1001284 (2011).
25. Blanchet, F. P. *et al.* Human immunodeficiency virus-1 inhibition of immunoamphisomes in dendritic cells impairs early innate and adaptive immune responses. *Immunity* **32,** 654–669 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We wish to thank members of the Molecular Virology laboratory for critical reading of the manuscript, N. Manel for $SIV3^+$ molecular clone and J. Luban for SIV delta Vpx and Vpx mutants. This work was supported by grants from the ERC (250333), Sidaction, ANRS and FRM 'équipe labéllisée' to M.B. N.L was supported by ANRS and SIDACTION fellowships; B.S. by ANRS fellowship; M.R. by CNRS/région Languedoc Roussillon fellowship. O.S. and N.C. are supported by grants from ANRS, Sidaction, ANR, European FP7 contract 201412 and Institut Pasteur.

Author Contributions M.B. and N.L. conceived the study and wrote the paper. M.B. and N.L. designed experiments and interpreted data. O.S. designed some experiments, interpreted data and edited the paper. N.L., B.S., N.C. and M.R. designed and performed experiments. C.C.-B. and E.S. provided technical assistance. N.L., B.S., N.C., M.B., A.Y., S.E. and O.S. discussed the data.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.B. (monsef.benkirane@igh.cnrs.fr) or N.L. (nadine.laguette@igh.cnrs.fr).

METHODS

Cell lines. Adherent and suspension cells were cultured in DMEM or RPMI supplemented with 10% fetal calf serum (FCS), ultraglutamine and antibiotics. All cell culture reagents were purchased from Lonza. Cell lines expressing Flag- and HA-tagged proteins were constructed using the previously described MMLV-based retroviral constructs[8,26] that contain a bicistronic transcriptional unit allowing for expression of a selectable marker (IL-2 receptor-α chain (pOZ-IL2Rα) or puromycin resistance gene (pOZ-puro)). IL2Rα-selected cell lines were selected using magnetic beads. Puromycin-selected cell lines were cultured in appropriate media supplemented with 1 μg ml^{-1} puromycin. shRNA-silenced cell lines were generated according to the manufacturer's instructions (Openbiosystem) and selected for resistance to puromycin. U937 and THP-1 cells were differentiated overnight with 30 ng ml^{-1} PMA (Sigma).

Plasmids. SIV3^{+} was a gift from N. Manel. HIV-LUC, VSV-G, MMLV packaging and A-MLV envelope have been previously described[27]. shRNA constructs were purchased from Openbiosystem. Vpx_{ROD}, $Vpx_{RCM\text{-}NG}$ and $Vpx_{RCM\text{-}GAB}$ were synthesized by MWG biotech. The SAMHD1 molecular clone was purchased from Invitrogen. All Vpx_{mac251}, Vpx_{ROD}, $Vpx_{RCM\text{-}NG}$ and $Vpx_{RCM\text{-}GAB}$ were subcloned in pOZ-IL2Rα expression vector with tags at the N terminus, and SAMHD1 and SAMHD1 mutants were subcloned in pOZ-puro expression vector with tags at the C terminus according to standard ligation procedures. SAMHD1(HD/AA) and SAMHD1-R mutants were generated using the Quickchange lightning kit (Agilent technologies) according the manufacturer's recommendations.

Virus production. Viral particles were produced from 293T cells using the standard phosphate calcium transfection protocol. Briefly, for HIV-LUC-G production, 293T cells were transfected with 8 μg HIV-LUC and 2 μg VSV-G encoding plasmid; for shRNA production, 293T cells were transfected with 4 μg shRNA construct, 4 μg packaging plasmid, 2 μg VSV-G encoding plasmid; for MLV transduction particles, 293T cells were transfected with 5 μg pOZ construct, 2.5 μg packaging plasmid and 2.5 μg A-MLV envelope encoding plasmid; and for VLP-Vpx production, 8 μg SIV3^{+} was co-transfected or not with 2 μg VSV-G encoding plasmid. Media was replaced 16 h after transfection and viruses were harvested 24 h later, filtered at 0.45 μm. When required, p24 concentration was measured by ELISA (Innogenetics).

Infection. Infection of THP-1, U937 and HeLa cells was performed by addition of 100, 50, 10, 5 or 1 ng p24 of HIV-LUC-G depending on the experiment. Viruses were added to cells and luciferase activity was measured 24 h after infection. VLP-Vpx treatment was performed for 2 h before infection by addition of RPMI-diluted VLP-Vpx to cells.

Cell extract preparation and western blot analysis. Whole-cell extracts were prepared with buffer containing 0.5% Triton X-100, 150 mM NaCl, 10 mM KCL, 1.5 mM $MgCl_2$, 0.5 mM EDTA 10 mM β-mercaptoethanol, 0.5 mM PMSF. Mouse anti-SAMHD1, anti-CUL4A and anti-DDB1 were purchased from Abcam. HA (11 clone 16B12) and tubulin antibodies were from Covance/Eurogentec and Sigma, respectively.

Immunopurification. Vpx_{mac251} was purified from 5×10^9 differentiated THP-1 cells stably expressing Flag- and HA-tagged Vpx_{mac251} (F/H-Vpx) by two-step affinity chromatography according to the standard method[8]. Extracts were first incubated with anti-Flag antibody conjugated agarose beads (Sigma), and the bound polypeptides were eluted with Flag peptide (Sigma) under native conditions. The Flag affinity-purified material was further immunopurified by affinity chromatography using anti-HA antibody conjugated agarose beads (Santa Cruz) and eluted under native conditions using HA peptide (Roche). Five per cent of immunoaffinity-purified F/H-Vpx or mock were resolved on SDS–PAGE and stained with the Silverquest kit (Invitrogen). The remainder of the eluate was stained with colloidal blue. Individual Coomassie-stained bands, or for closely migrating bands, regions of the gel were excised and subsequently analysed by tandem mass spectrometry at the Harvard Medical School Taplin Biological Mass Spectrometry facility.

Quantification of HIV-1 total DNA. Before infection, viral stocks were treated for 1 h at 37 °C with 100 U ml^{-1} of DNaseI (Roche). 3×10^5 THP-1-scr or THP-1-shSAMHD1 cells were infected with 100 ng of HIV-LUC-G or with heat-inactivated HIV-LUC-G. Cells were washed twice in PBS 2 h after infection, harvested 22 h later, washed twice in PBS, and DNA was extracted using the QIAamp Blood DNA Minikit (Qiagen). Quantification of viral DNA was performed by quantitative PCR (qPCR) using previously described probes[27] allowing for amplification of the luciferase gene on a LightCycler 480 system (Roche). Viral DNA was normalized using 7SK-specific probes.

Preparation of MDDCs and MDMs. Human buffy coats were obtained from Etablissement Français du Sang (EFS). Monocytes were purified from total peripheral blood mononuclear cells after Ficoll gradient separation with CD14 MicroBeads (Miltenyi Biotec) or using the standard adhesion protocol. Human monocyte-derived dendritic cells (MDDCs) were generated by incubation of CD14 purified monocytes in IMDM medium supplemented with 10% FCS, 2 mM L-glutamine, 100 IU ml^{-1} penicillin, 100 mg ml^{-1} streptomycin, 10 mM HEPES, 1% non-essential amino acids, 1 mM sodium pyruvate, 10 ng ml^{-1} GM-CSF and 50 ng ml^{-1} IL-4 (Miltenyi Biotec). On day 4, two-thirds of the culture medium was replaced by fresh medium containing GM-CSF and IL-4. Immature MDDCs (iDCs) were harvested and further used at day 6.

MDMs were generated by 7 days of stimulation with 50 ng ml^{-1} of recombinant human granulocyte-macrophage colony-stimulating factor (rhGM-CSF) (Immunotools).

Efficient differentiation of MDM and MDDCs was verified by flow cytometry using antibodies against CD1a, CD14, CD80, CD83 (BD Bioscience) and CD86 (B72 clone, Invitrogen). MDDCs were more than 98% DC-SIGN positive.

siRNA transfection of MDDCs. scrambled siRNA, siSAMHD1-1 (5′-GAUUCA UUGUGGCCAUAUA-3′) and siSAMHD1-3 (5′-CAACCAGAGCUGCAGAU AA-3′) were synthesized by MWGoperon. siDYN2 was from Qiagen. HiPerFect Reagent (Qiagen) was used for transfection in accordance with the manufacturer's recommendations. In brief, 5×10^5 iDCs were transfected with 100 nM siRNA in 500 ml of IMDM/1% FCS medium in 12-well plates. A second round of transfection was performed 24 h later. Specific gene knockdowns were assessed by immunoblot.

Transduction of MDDCs. One day after the second round of silencing, MDDCs were transduced with a lentiviral vector expressing the GFP or luciferase gene (LV-GFP and LV-LUC, respectively) or VSV-G-pseudotyped HIV (HIV-G). Briefly, 2.5×10^5 MDDCs were seeded in 24-well plates and exposed to the indicated doses of lentiviral vectors or HIV-G (from 1 to 100 ng p24 per ml). After overnight incubation, the medium was replaced with fresh medium containing GM-CSF and IL-4. Four days after transduction, GFP or p24 expression was analysed by flow cytometry using a FacsCalibur (Becton Dickinson) with FlowJo software or luciferase activity was measured using a Mithras luminometer (Berthold technologies).

26. Kumar, D., Shadrach, J. L., Wagers, A. J. & Lassar, A. B. Id3 is a direct transcriptional target of Pax7 in quiescent satellite cells. *Mol. Biol. Cell* **20,** 3170–3177 (2009).
27. Sobhian, B. *et al.* HIV-1 Tat assembles a multifunctional transcription elongation complex and stably associates with the 7SK snRNP. *Mol. Cell* **38,** 439–451 (2010).

ARTICLE

doi:10.1038/nature10143

Role of the ubiquitin-like protein Hub1 in splice-site usage and alternative splicing

Shravan Kumar Mishra[1], Tim Ammon[1], Grzegorz M. Popowicz[2], Marcin Krajewski[2]†, Roland J. Nagel[3]†, Manuel Ares Jr[3], Tad A. Holak[2] & Stefan Jentsch[1]

Alternative splicing of pre-messenger RNAs diversifies gene products in eukaryotes and is guided by factors that enable spliceosomes to recognize particular splice sites. Here we report that alternative splicing of *Saccharomyces cerevisiae SRC1* pre-mRNA is promoted by the conserved ubiquitin-like protein Hub1. Structural and biochemical data show that Hub1 binds non-covalently to a conserved element termed HIND, which is present in the spliceosomal protein Snu66 in yeast and mammals, and Prp38 in plants. Hub1 binding mildly alters spliceosomal protein interactions and barely affects general splicing in *S. cerevisiae*. However, spliceosomes that lack Hub1, or are defective in Hub1–HIND interaction, cannot use certain non-canonical 5′ splice sites and are defective in alternative *SRC1* splicing. Hub1 confers alternative splicing not only when bound to HIND, but also when experimentally fused to Snu66, Prp38, or even the core splicing factor Prp8. Our study indicates a novel mechanism for splice site utilization that is guided by non-covalent modification of the spliceosome by an unconventional ubiquitin-like modifier.

Covalent modification of proteins by ubiquitin and related proteins (collectively called ubiquitin-like modifiers, UBLs) often critically alters substrate activity by influencing metabolic stability, binding behaviour or localization[1]. The switch-like properties of UBLs are crucial for pathways that regulate, for example, signal transduction, protein sorting, DNA repair, and development[1]. Covalent conjugation of a UBL to a substrate's target residue is ATP dependent, involves an enzyme cascade, and usually requires a free di-glycine (GG) motif at the protruding carboxy-terminal end of the UBL. Archetypal UBLs (ubiquitin, SUMO, Rub1 (also known as Nedd8)) are expressed as inactive precursors with C-terminal extensions. These extensions are removed by UBL-specific proteases, exposing the crucial C-terminal GG motif. Enzymes of this class also mediate UBL deconjugation, thus making the UBL-dependent switch reversible[1].

Hub1 (homologous to ubiquitin; known as UBL5 or beacon in mammals), another evolutionarily highly conserved UBL, is unique in lacking a protruding C-terminal tail with a GG motif. Instead, Hub1 possesses a C-terminal double tyrosine (YY) motif, followed by a non-conserved amino acid residue[2,3]. Although Hub1 from various organisms has been studied to some extent[4–8], its function remains poorly understood. Whereas *S. cerevisiae* cells deficient in Hub1 are viable and exhibit only minor phenotypes under normal growth conditions[6,7], the corresponding mutant of *Schizosaccharomyces pombe* is lethal[4,8]. One study reported that Hub1 forms covalent conjugates similar to ubiquitin and proposed that Hub1 is synthesized as a precursor and matured by processing C terminally of the YY motif[6]. However, no Hub1-specific processing, conjugation or deconjugation enzymes have been identified. Further studies have ruled out that Hub1 functions as a covalent protein modifier[7,8]; in fact, Hub1 was found to bind proteins non-covalently and independently of ATP, and the YY motif was shown to be nonessential[7,8].

Hub1 has been linked to various physiological functions, including cell cycle progression and polarized growth[6], the mitochondrial unfolded protein response[9], and mRNA splicing[4,8]. Conditional mutants of *S. pombe HUB1* show moderate RNA splicing defects, particularly at high temperatures, and Hub1 formed a non-covalent association with the spliceosomal (U4/U6.U5) tri-small nuclear ribonucleoprotein particle (snRNP) protein Snu66 (refs 4, 8). However, how the Hub1–Snu66 interaction affects splicing is unclear. It has been proposed that Hub1 is required for the nuclear localization of Snu66 (ref. 4), but Hub1 may affect the spliceosome directly and influence its activity.

Here we show that Hub1, through binding to Snu66, modifies the spliceosome in a way that enables it to tolerate and use certain non-canonical 5′ splice sites. We discovered that Hub1 binds Snu66 through an element called HIND in a unique, sequence-specific manner. We propose that Hub1 operationally resembles UBLs, with the important difference that Hub1 modifies substrates through non-covalent binding.

Hub1 binds to HINDs of spliceosomal proteins

Hub1 has been shown to bind the tri-snRNP protein Snu66 in yeast two-hybrid (Y2H) assays[4,10]. To verify this interaction *in vivo*, we raised antibodies specific for yeast Hub1 and Snu66 for co-immunoprecipitation assays. The antibodies against Hub1 bring down Snu66 from cell extracts and vice versa (Fig. 1a, b). Similar experiments with an altered form of Hub1 (YY changed to AA) showed that binding to Snu66 was independent of the YY motif (data not shown).

Amino-terminal fragments of Snu66 were both sufficient and required for Hub1 binding. Inspection of the protein sequence revealed two highly similar elements (72% identity) arranged in tandem (Fig. 1c, d). Fragments harbouring these repeats, either singly or in tandem, strongly interacted with Hub1 in Y2H assays (Supplementary Fig. 1a). Notably, the elements do not seem to bind ubiquitin or SUMO in Y2H assays, and Hub1 does not seem to bind classical ubiquitin-binding motifs like UBA or UIM (data not shown). The minimal polypeptide sequence defined by this assay was 18–19 amino acids long; because it had no obvious similarity to known motifs, we termed it HIND

[1]Department of Molecular Cell Biology, Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany. [2]NMR Spectroscopy, Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany. [3]Center for Molecular Biology of RNA, Department of Molecular, Cell & Developmental Biology, University of California, Santa Cruz, California 95064, USA. †Present addresses: Institut für Biomedizinische Technik, ETH Zürich, Wolfgang-Pauli-Str. 10, 8093 Zürich, Switzerland (M.K.); Enzymology Research and Development, Life Technologies, 850 Lincoln Centre Drive, Foster City, California 94404, USA (R.J.N.).

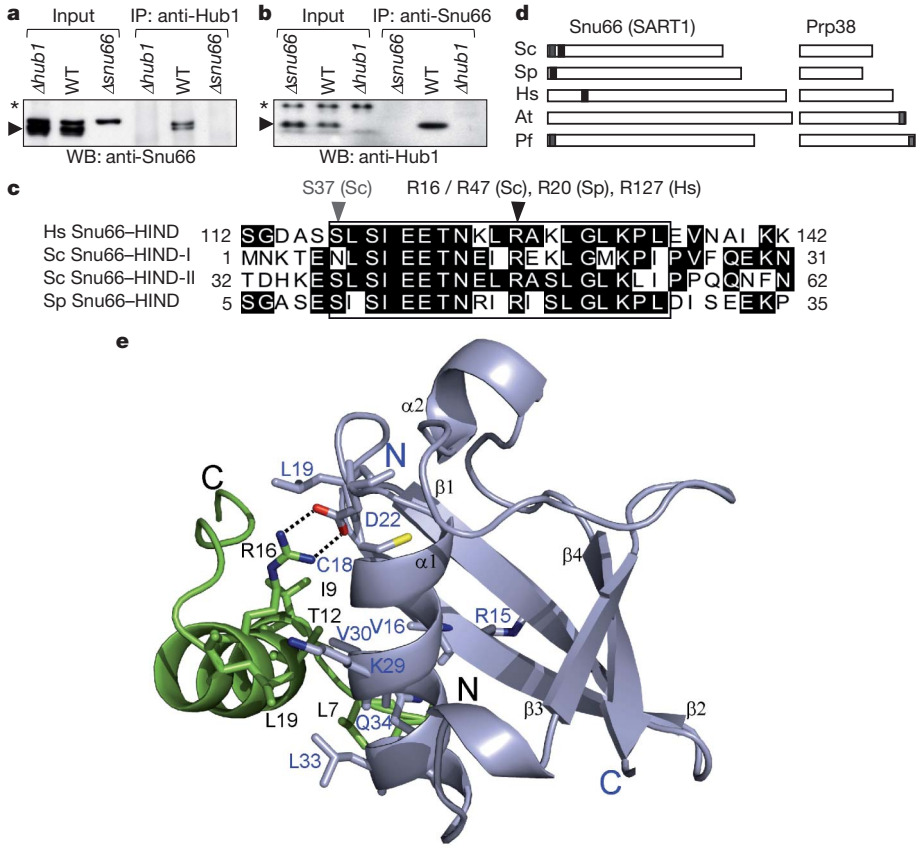


Figure 1 | Hub1 binds to HIND. **a**, Immunoprecipitation (IP) of Hub1 and co-immunoprecipitation of Snu66 from the indicated yeast strains. WB, western blot; WT, wild type. **b**, Immunoprecipitation of Snu66 and co-immunoprecipitation of Hub1 from the indicated yeast strains (asterisks denote cross-reacting signals). **c**, Clustal-W alignment of Snu66–HINDs from different organisms. Numbers give the positions of amino acids, and arrows indicate conserved Arg and Ser residues. At, *Arabidopsis thaliana*; Hs, *Homo sapiens*; Pf, *Plasmodium falciparum*; Sc, *Saccharomyces cerevisiae*; Sp, *Schizosaccharomyces pombe*. **d**, Diagram of homologues of the spliceosomal proteins Snu66 (SART1) and Prp38, and the occurrence of HINDs (shaded). **e**, Crystal structure of Hub1 in complex with HIND-I peptide shown as ribbon plot. The interaction interface is formed by a salt bridge between Arg 16(47) (numbers in parentheses refer to HIND-II) of Snu66 and Asp 22 of Hub1, accompanied by a patch of hydrophobic interactions. The hydrophobic residues of HINDs that participate in the Hub1-binding interface are Leu 7(38), Ile 9(40), Thr 12(43), Ile 15 (Leu 46), Arg 16(47) (Cβ and Cγ), Leu 19(50), Met 21 (Leu 52) and Ile 24(55). On Hub1, the interface is formed by Leu 19, Val 30, and additionally Met 1, Val 16, Lys 17 (Cβ, Cγ, Cδ), Cys 18, Lys 29 (Cβ, Cγ, Cδ) and Leu 33.

(Hub1-interaction domain). GST pull-down assays with purified proteins confirmed that Hub1–Snu66 binding is direct and mediated via HIND (Supplementary Fig. 1b, c). Gel filtration analysis indicates that the N-terminal domain of Snu66 can bind two Hub1 molecules (Supplementary Fig. 1d).

Snu66 proteins (also known as SART1) from Saccharomycotina (for example, *S. cerevisiae*, *Candida*) and some Stramenopiles possess two N-terminal HINDs, but the homologues from *S. pombe* and vertebrates possess only one HIND at this position (Fig. 1d). As Hub1 binds Snu66 via its HIND in *S. pombe* and in humans also (Supplementary Fig. 2a–e), the mechanism of Hub1 recruitment seems to be conserved. Intriguingly, plant Snu66 homologues (and also for example, Amoebazoa) lack HIND sequences; this absence is compensated by HINDs found in C-terminal extra domains of proteins related to the spliceosomal protein Prp38 (Fig. 1d and Supplementary Fig. 2f). Furthermore, in *Plasmodium*, functional HINDs are present in both Snu66 and Prp38 homologues, as they bind Hub1 and involve the same surface interface (Supplementary Fig. 2g, h). Because in yeast and mammalian cells Snu66 and Prp38 are constituents of the spliceosome[11], HINDs may be evolutionarily associated (and perhaps restricted) to splicing proteins. Moreover, as the identity of the HIND-bearing spliceosomal protein seems to be flexible, HINDs may function irrespective of their exact positioning within the spliceosome (see later). By extension, this observation indicates that Hub1 may affect spliceosome activity as a whole, rather than functionally modulate the protein it directly binds. As Hub1 colocalizes with Snu66 in nuclear speckles in human cells (Supplementary Fig. 3a), it seems likely that the function of Hub1 in splicing is conserved.

Structure of Hub1 in complex with HIND

To determine whether Hub1 functions as a non-covalently acting modifier by interacting with HIND, we sought to gain molecular insights into the Hub1–HIND interaction. The characterization of peptides corresponding to the two HINDs of *S. cerevisiae* Snu66 (HIND-I, 18 amino acids; HIND-II, 19 amino acids) by circular dichroism (CD) and NMR revealed that an isolated HIND peptide is apparently helical in solution (Supplementary Fig. 4a, b), an unusual feature for such short peptides.

Next, we solved the structure (1.4 Å) of Hub1 in complex with each Snu66 HIND peptide. The structure of Hub1 is highly similar to ubiquitin[2,3,12] (PDB code 1UBI; root mean squared deviation (r.m.s.d.) of 0.88 Å for the main chain heavy atoms) and SUMO[13] (PDB code 1A5R; r.m.s.d. of 1.72 Å), and consists of a half-open β barrel completed with two flanking α helices; the secondary structure elements have a ββαβαβ pattern (Fig. 1e, Supplementary Figs 5, 6 and Supplementary Table 2).

The observed interactions of Hub1 with the Snu66 HIND-I and HIND-II peptides are almost identical (Fig. 1e and Supplementary Fig. 5a). The bound peptide forms an 11-residue helix (Ile 3–Leu 13,

numbering of HIND-I) with its C-terminal part flipped over along the peptide helix (Fig. 1e and Supplementary Fig. 5a). The whole Hub1–HIND interface has a surface of about 500 Å^2. On HIND binding, the main-chain fold of Hub1 does not change, but a number of side chains are significantly affected (Supplementary Fig. 5b). The Hub1–Snu66 interactions seen in the X-ray structures were fully corroborated by our mutational studies. For example, replacement of the salt-bridge-forming residues Asp 22 of Hub1 and Arg 16(47) (number in parentheses refers to residues in HIND-II) of Snu66 by Ala abolished Hub1–Snu66 binding (Supplementary Fig. 5c–f).

Most unexpectedly, the structure of Hub1–HIND revealed a new binding paradigm unseen in interactions of ubiquitin and ubiquitin-like proteins with their binding partners. Most ubiquitin-binding modules of ubiquitin receptors bind to the hydrophobic surface of ubiquitin centred on Ile 44 (ref. 14), which is almost exactly on the site opposing the HIND-binding face of the ubiquitin fold (Supplementary Fig. 6a). The ubiquitin interactions with ubiquitin receptors have mostly hydrophobic character and are usually weak ($K_d \geq 100$ μM), whereas the Hub1–HIND interaction comprises a strong salt bridge accompanied by several hydrophobic contacts and high affinity (K_d Hub1–Snu66: 0.59 ± 0.07 μM; K_d Hub1–HIND-I: 1.69 ± 0.27 μM). The mode of interaction of Hub1–HIND is also clearly different from that seen for SUMO interacting with SUMO-interaction motifs (SIM)[15], although SIMs bind to a similar surface of the ubiquitin fold (Supplementary Fig. 6b).

By performing heteronuclear single quantum coherence (HSQC) NMR, we found that $^{15}N^2H$-labelled Snu66, including its HIND-bearing N-terminal domain, is unstructured (Supplementary Fig. 7a, b). As co-expression of Hub1 significantly increased the solubility of Snu66 in bacteria (Supplementary Fig. 7c), we speculated that Hub1 might influence Snu66 folding. Indeed, when unlabelled Hub1 was titrated into $^{15}N^2H$-labelled Snu66, several distinct peaks appeared, indicating that a part of the protein acquires structure after Hub1 binding (Supplementary Fig. 7a, b). Comparison of an N-terminal 65-amino-acid fragment and full-length Snu66 revealed that Hub1-induced folding is restricted to the HIND-containing N terminus of Snu66. This finding further corroborates the idea that Hub1 modifies the spliceosome rather than modulating the properties of an individual binding partner.

Hub1 modifies the spliceosome

Snu66 is a conserved component of the spliceosomal tri-snRNP complex[16,17]. To investigate whether Hub1 alters the composition of the spliceosome, we immunoprecipitated Snu66 from wild-type, *Δhub1* or *HUB1*-overexpressing strains (*Δsnu66* strain as control), and identified co-purifying proteins by Orbitrap mass spectrometry (Supplementary Fig. 8a–c). The set of identified proteins in all three samples significantly overlapped with the catalogue of proteins in the yeast spliceosomal complex B (penta-snRNP)[18,19]. Although we did not observe major changes in the protein composition of the three samples, in Hub1-deficient cells we reproducibly found an overrepresentation of certain proteins from U1 and U2 snRNPs, but not of the tri-snRNP (Supplementary Fig. 8a–d); by contrast, no significant spliceosomal alterations were observed in the strain overexpressing Hub1. We confirmed these findings by Snu66-directed immunoprecipitation using strains that express tagged versions of selected proteins from the tri-, U1-, and U2-snRNPs from their genomic loci (Supplementary Fig. 8d). We conclude that modification of the spliceosome by Hub1 only moderately affects spliceosome composition, so that the basic makeup of the snRNPs is preserved. Hub1 does bind spliceosomes *in vivo*, as we could co-immunoprecipitate the central spliceosomal protein Prp8 with Hub1 antibodies, in a reaction that is mediated by Snu66 (Supplementary Fig. 8e).

Functional links to the spliceosome

Hub1 is present in both cytosol and nucleus in yeasts and mammals, whereas Snu66 appears to be mostly nuclear (Supplementary Fig. 3a, b)[4,8]. It was reported that Hub1 is required for the nuclear localization of Snu66 (ref. 4), but we found that Snu66 is nuclear even in *S cerevisiae Δhub1* cells, and that a Snu66 variant deficient in Hub1 binding is nuclear in *S. cerevisiae*, *S. pombe* and human cells (Supplementary Fig. 3a, b and data not shown). Although in *S. pombe HUB1* and *SNU66* are essential[4,8], mutants deficient in Hub1–Snu66 interaction are viable (Supplementary Fig. 10a, b). This finding, together with the detection of significant cytosolic pools of Hub1 and the observation that Hub1 is apparently more abundant in cells than Snu66 (data not shown), indicates that the function of the Hub1 modifier may not be restricted to splicing.

We observed synthetic sickness or lethality of *Δhub1* and *Δsnu66* mutants if they were combined with mutant alleles of a number of spliceosomal genes (Supplementary Fig. 9a, b). Notably, the phenotype of *Δhub1 prp8** (*prp8** refers to *prp8(P1384L)*) could not be rescued by expressing *hub1(D22A)*, which encodes a Hub1 variant defective in Snu66 interaction (Fig. 2a). Similarly, the lethal *Δsnu66 prp8** double mutant could not be complemented by expression of Snu66 variants that lack the two HINDs (Fig. 2b). Thus, we conclude that Hub1, Snu66 and the Hub1–HIND interaction are indeed relevant for splicing.

As noted earlier, HIND elements in *S. cerevisiae* are located within the N-terminal domain of Snu66, but in plants, HIND is present as a C-terminal extension of the spliceosomal protein Prp38 (Fig. 1d). To address whether the specific localization of HIND is important, we fused a HIND element to the C terminus of yeast Prp38 (*PRP38–HIND*) and determined functionality as above. We found that *PRP38–HIND* could indeed functionally rescue *snu66(ΔHIND)* in the synthetic lethality assay with *prp8**, and that this activity depends on HIND residues crucial for Hub1 binding (Supplementary Fig. 9d). This finding confirms a remarkable plasticity of the Hub1 modifier, in that the exact positioning of HIND (and thereby of Hub1), on the spliceosome is not critical for function.

The *S. cerevisiae Δhub1 Δsnu66* double—but not the corresponding single—mutant is temperature sensitive (Supplementary Fig. 9b). This non-epistatic behaviour suggests partially separate functions and may again point to roles of Hub1 in addition to splicing. Whereas the removal of the YY motif does not interfere with Hub1 activity[4,7,8], Hub1 variants possessing certain charged C-terminal extensions failed to support the vital activity of Hub1 in *S. pombe* (Supplementary Fig. 10c), although the proteins can still bind HIND (Supplementary Fig. 10d). Moreover, also in *S. cerevisiae*, expression of a Hub1 variant with a short extension (*hub1–DD*) could not suppress the synthetic sickness of *Δhub1 prp8** (Fig. 2a). Thus, in addition to the HIND-binding surface the area neighbouring the C terminus of Hub1 is functionally important, perhaps for additional physical interactions.

To investigate the splicing competence of Hub1-deficient cells directly, we used a splicing-sensitive microarray that can distinguish spliced from unspliced RNAs of almost all intron-containing *S. cerevisiae* genes[20,21]. Onto this array we hybridized RNA samples from wild-type, *Δhub1*, *Δsnu66* and *Δhub1 Δsnu66* cultures (Fig. 2c). We found virtually no splicing defects in the *Δhub1* mutant for all transcripts, with the only discernable exception of *RPL34B*, the splicing of which was mildly affected. On the other hand, we noticed small splicing defects for *Δsnu66*, which were aggravated if the cells were additionally defective in Hub1 (*Δhub1 Δsnu66*) (Fig. 2c and Supplementary Table 3). To confirm these data, we isolated RNA from these strains and analysed splicing of a selected set of RNAs by quantitative polymerase chain reaction with reverse transcription (RT–qPCR). Again, splicing was virtually normal in *Δhub1* cells, and only a small splicing defect was discernable for *RPL34B* (Supplementary Fig. 11a). However, this small defect in *RPL34B* splicing is rather unlikely to affect the level of this ribosomal protein, as splicing of *RPL34A*, the gene product of which is almost identical to that of *RPL34B*, is not affected (data not shown). We found that *HUB1* transcripts are inducible by cadmium and on oxidative stress (Supplementary Fig. 11b),

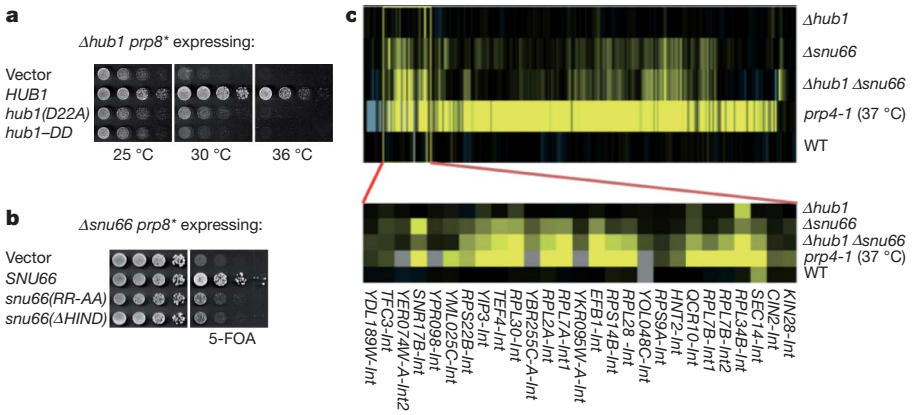


Figure 2 | Functional links to splicing. **a**, Rescue of synthetic sickness of *Δhub1 prp8** by expression of *HUB1* wild type, but not *hub1(D22A)* and *hub1–DD*. **b**, Rescue of synthetic lethality of *Δsnu66 prp8** by expression of *SNU66* wild type, but not by *snu66(RR-AA)* or *snu66(ΔHIND)*. Wild-type *SNU66* expressed from a *URA3*-containing plasmid was shuffled out by counter-selection on 5-fluoroorotic acid (5-FOA) plates. **c**, Microarray analysis for splicing defects. Diagram of intron accumulation (yellow) in strains deleted of *HUB1* and *SNU66* (data in Supplementary Table 3). The top panel shows relative splicing efficiency for almost all yeast introns as measured by splicing sensitive microarrays[20,21]. In each case RNA is derived from the indicated mutant, and wild-type RNA (comparing two different RNA preparations from wild-type cells) is used for reference. The temperature-sensitive splicing mutant *prp4-1* is the positive control. The bottom panel is an expanded view of a set of transcripts that are particularly affected in *Δhub1 Δsnu66* and also includes an intron of *RPL34B* that is mildly affected in *Δhub1*.

but we detected no significant *Δhub1*-specific splicing defects under these (or heat shock) conditions (Supplementary Fig. 11c, d). Thus, Hub1 does not seem to significantly affect general splicing, and splicing defects only occur if the spliceosome has additional deficiencies. This indicates that Hub1 might have a silent (and redundant) role in general splicing, or that Hub1 fulfils a specific, splicing-related function.

Splice-site usage and alternative splicing

Splice sites in *S. cerevisiae* show very little sequence variation, and alterations in the sequence can affect splicing[22]. To address whether Hub1 affects splice-site usage, we used a reporter assay based on an intron-containing *RP51*–LacZ* fusion[23] (Fig. 3a) monitored by both RT–PCR and β-galactosidase (LacZ) activity. For the consensus 5′ splice site (GUAUGU), again we observed no splicing defects for *Δhub1*, but defects for *Δsnu66*, *prp8** and *Δhub1 prp8** (Fig. 3b). Intriguingly, when we altered the 5′-splice-site sequence in the reporter construct, some 5′-splice-site variants were similarly used in both wild-type and *Δhub1* cells, but certain others required specifically the presence of Hub1 for proper splicing (Fig. 3a, b and data not shown). In fact, splicing via the mutant 5′ splice site GUAUAU was nearly as defective in *Δhub1* as in *prp8**, and was almost abolished in *Δhub1 prp8** (Fig. 3b). Importantly, this splicing defect also occurred in strains that only express a Hub1 defective in Snu66 binding (*hub1(D22A)*), or Hub1 harbouring an abnormal C terminus (*hub1–DD*) (Fig. 3c). To corroborate and extend these findings we also used another splicing sensitive *ACT1–CUP1* reporter assay[24]. Again we found splicing defects of *Δhub1* cells for certain 5′-splice-site alterations (GUAUAU, GUCUGU), but not significantly for variants of the 3′ splice site (G/AG, U/UG instead of U/AG; slash symbolizes the intron/exon boundary) or the branch point (UCCUAAC, UACUACC instead of UACUAAC) (Supplementary Fig. 12).

Most metazoans exhibit a high variation of 5′ splice sites, and this plasticity is exploited for alternative splicing[22,25]. Regulation of these events seems to be guided by factors that enable the spliceosome to recognize divergent 5′ splice sites. In *S. cerevisiae*, in which 5′ splice sites are nearly invariant, alternative splicing is extremely rare[22]. In this organism, *SRC1* is the only known case of alternative splicing leading to two different proteins with different functions[26–28]. In both cases, a single (130- or 126-nucleotide long) intron is excised, yet by using two different, overlapping 5′ splice sites. Notably, both 5′ splice sites differ significantly from the consensus (GCAAGU, GUGAGU).

SRC1 was present on the splicing array we used, but the two splice forms could not be individually identified because of their overlapping

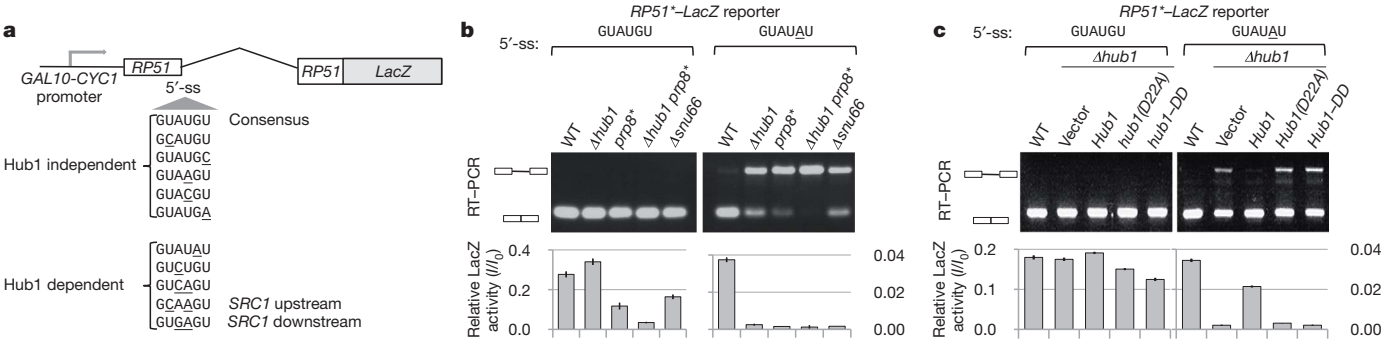


Figure 3 | Splice-site usage. **a**, Scheme for the requirement of Hub1 for efficient utilization of canonical and non-canonical 5′ splice sites (5′-ss; underlines mark changes) inserted into *RP51*–LacZ* fusions[23]. The individual 5′ splice sites of *SRC1* in the context of the reporter assay were used. **b**, **c**, Splicing defects in wild type and mutants with canonical (GUAUGU) and non-canonical (GUAUAU) 5′ splice sites. RT–PCR assay of total RNA isolated from yeast expressing different version of intron-containing *RP51*–LacZ* fusions. Bottom panels show ratios of corresponding LacZ (β-galactosidase) activities (*I*) to the activity from an intron-less (I_0) construct. Error bars show standard deviation of three samples. Strains in **c** are *Δhub1* and complemented by *HUB1* alleles.

5′ splice sites. When we sequenced the complementary DNA across the exon/exon boundaries of spliced *SRC1*, however, we observed a sequence mixture corresponding to the two splice forms in wild-type cells (and also in *Δsnu66* and *prp8** mutants), but largely only one form (corresponding to splicing via the downstream 5′ splice site) in *Δhub1* mutants (Fig. 4a and Supplementary Fig. 13a, b). To corroborate this finding, we made use of a chromosomally expressed TAP-tagged Src1 variant monitoring the expression of the two splice variants[28]. Indeed, splicing via the upstream 5′ splice site (GCAAGU) was again almost completely absent in Hub1-deficient cells, whereas splicing via the downstream 5′ splice site was normal or even more pronounced (Fig. 4b). Moreover, splicing via the upstream 5′ splice site was equally defective in cells that express Hub1 defective in HIND interaction (*hub1(D22A)*), whereas both forms were reduced in *Δsnu66* (Fig. 4b). We also detected both forms of the Src1 protein in *Δprp17* and *prp8** mutants (Supplementary Fig. 13c), indicating that the choice of *SRC1* 5′ splice site is not significantly altered in these general splicing mutants. Notably, we could make *SRC1* splicing via the upstream 5′ splice site independent of Hub1 by mutating this 5′ splice site, but this caused a repression of splicing via the downstream 5′ splice site (Supplementary Fig. 13d). Conversely, certain mutations in the downstream 5′ splice site strongly affected splicing via the upstream 5′ splice site even when Hub1 was present (wild-type cells). We also tested the two *SRC1* 5′ splice sites individually in the *RP51**–*LacZ* reporter and found that in isolation, both non-canonical elements require Hub1 for full splicing activity (Fig. 3a; data not shown). This indicates that the characteristic differential Hub1 dependence of *SRC1* alternative splicing requires the tandem arrangement of overlapping 5′ splice sites. In this context it is interesting to note that alternative *SRC1* splicing might be regulated, as we found that specifically the Hub1-dependent splicing product is enriched in the G2 phase of the cell cycle (Supplementary Fig. 13e).

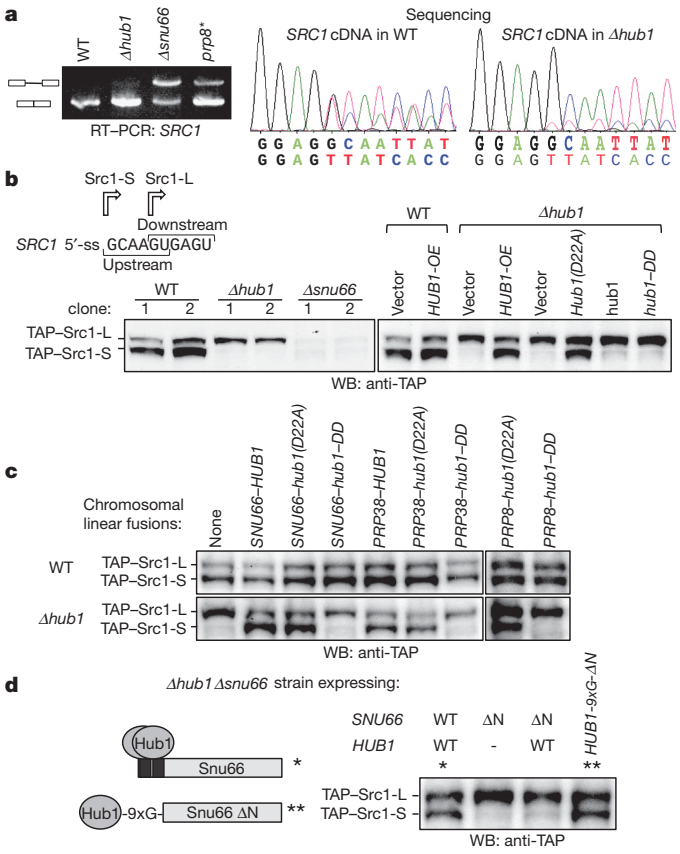


Figure 4 | Alternative splicing of *SRC1*. **a**, RT–PCR assay of *SRC1* transcripts in different mutants (left panel), and sequencing of *SRC1* cDNA from wild-type (WT) and *Δhub1* strains (right panels). Only general splicing defects of *Δsnu66* and *prp8** mutants are detectable by RT–PCR as the 5′ splice sites of *SRC1* are overlapping, but alternative splicing defects via the upstream 5′ splice site in the *Δhub1* strain are revealed by sequencing. **b**, Chromosomal *SRC1* was N-terminally TAP tagged to detect the two gene products (TAP–Src1-L, TAP–Src1-S) generated by alternative splicing. Splicing via the upstream (GCAAGU) and downstream (GUGAGU) 5′ splice sites (5′-ss) is indicated (top left). Alternative splicing defects (usage of the upstream 5′ splice site) in *Δhub1* and general splicing defects in *Δsnu66* were detected by western blotting using TAP-tag-specific antibodies. Wild-type and mutant strains, and *Δhub1* strains expressing Hub1 variants, or which overexpress *HUB1*, were used. **c**, *SRC1* alternative splicing is restored in *Δhub1* harbouring chromosomally expressed linear fusions of *PRP8*, *PRP38* and *SNU66* genes with *HUB1* or *hub1(D22A)* but not with *hub1–DD* (assay similar to **b**). **d**, *SRC1* alternative splicing is supported by a linear fusion of Hub1 with Snu66ΔN in a *Δhub1 Δsnu66* strain. The *Δhub1 Δsnu66* strain with TAP-tagged Src1 was transformed with plasmids expressing *HUB1*, *SNU66*, or a construct encoding Hub1 fused to the N terminus of the Snu66 variant that lacks the HIND region (ΔN), separated by a poly-glycine (9×G) linker. Asterisks indicate the two ways of Hub1–Snu66 complex formation.

Compared to *S. cerevisiae*, *S. pombe* possesses a much higher number of intron-containing genes, and its splicing machinery resembles more closely its mammalian counterpart[22,29]. Similarly, the sequences of *S. pombe* 5′ splice sites are highly divergent and several cases of alternative splicing are known[22,30]. To address whether Hub1 also mediates selective splicing in *S. pombe* we focused on two different pre-mRNAs. We tested *CDC2* (encoding cyclin-dependent kinase), as it possesses four introns with different 5′ splice sites. We found that splicing of intron 3 in particular, which possesses a rarely used 5′ splice site (GUAUAU; Hub1-dependent in *S. cerevisiae*; Fig. 3a), needs Hub1 for normal splicing (Supplementary Fig. 14a–d). *ZAS1* (encoding a transcription factor), on the other hand, is known to be alternatively spliced as sometimes intron 2 is retained, yielding a larger open reading frame[30]. Again we found that Hub1 is needed for normal *ZAS1* splicing and that intron inclusion occurs in *hub1* mutants at a much higher frequency (Supplementary Fig. 14d). Thus, Hub1-guided selective splicing is a conserved mechanism, and may occur in all eukaryotes.

Using the *SRC1* splicing assay, we examined the significance of our initial observation that HINDs are present on Snu66, Prp38 or both spliceosomal proteins in different organisms. Remarkably, alternative *SRC1* splicing was normal in *Δhub1 S. cerevisiae* cells in which linear (C-terminal) fusions of either Snu66 or Prp38 with Hub1 were chromosomally expressed (Fig. 4c). *SRC1* splicing was defective when the Hub1 moiety of the fusions was C-terminally altered (Hub1–DD), but fully functional when the Hub1 moiety was deficient in Snu66 interaction. Hub1 fused to Snu66ΔN (lacking the HINDs) via a polyglycine linker may also support Hub1-dependent alternative *SRC1* splicing even if free Hub1 and wild-type Snu66 are absent (Fig. 4d). Surprisingly, when we fused HIND-binding defective Hub1 (Hub1(D22A)) to the core spliceosomal protein Prp8, alternative *SRC1* splicing was supported (Fig. 4c). These findings indicate that linear Hub1 fusions functionally mimic HIND-mediated Hub1 recruitments, and show that Hub1-dependent effects on splicing act independently of a specific localization on the spliceosome.

Conclusions

Alternative splicing substantially increases the gene product repertoire and is a major source of cell type differentiation[22,25]. Metazoans in particular use this mechanism extensively, and it is estimated that the majority of human pre-mRNAs undergo alternative splicing[22]. Conventional alternative splicing is largely controlled by positively acting SR (Ser-Arg) proteins and negatively acting heterogeneous nuclear ribonucleoprotein particles (hnRNPs).

We have discovered a new principle for splice-site utilization, which involves non-covalent binding of the ubiquitin-related protein Hub1 to the spliceosome. Notably, Hub1 in *S. cerevisiae* is not required for general splicing and the usage of canonical 5′ splice sites

(Fig. 2c), but is required for the usage of certain non-canonical 5′ splice sites. Hub1 might function as a 'plasticity factor', which, by relaxing the spliceosome's specificity, enables it to act productively on divergent 5′ splice sites. Given the high conservation of Hub1 and HIND, and the fact that Hub1-dependent splicing operates through evolutionary divergent spliceosomes, it seems highly likely that Hub1-controlled splicing occurs universally in eukaryotes. SR proteins and hnRNPs involved in spliceosome targeting do not seem to exist in *S. cerevisiae*[22,31], and thus the Hub1-dependent mechanism may be evolutionarily older. It seems plausible that the more elaborate SR/hnRNP-guided mechanism may have co-evolved with the rise in gene complexity to generate multiple different gene products from a single gene.

Hub1 is structurally very similar to ubiquitin and equally highly conserved and ancient, yet the two proteins function in completely different ways. Hub1 binds proteins only non-covalently, and at least for splicing, Hub1 seems to functionally alter the complex it binds to, rather than modulating its direct binding partner. Because of this distinctive property, it is attractive to speculate that Hub1 might alter the structure of the spliceosome or provide novel binding surfaces for physical interactions.

METHODS SUMMARY

Yeast strains are listed in Supplementary Table 1. Protein methods, mass spectrometric analysis, interaction studies, splicing assays and microscopy are standard techniques and are described in Methods. NMR spectroscopy and X-ray crystallography are detailed in Methods, and the data collection and refinement statistics are summarized in Supplementary Table 2.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 4 November 2010; accepted 20 April 2011.
Published online 25 May 2011.

1. Hochstrasser, M. Origin and function of ubiquitin-like proteins. *Nature* **458,** 422–429 (2009).
2. McNally, T. *et al.* Structural analysis of UBL5, a novel ubiquitin-like modifier. *Protein Sci.* **12,** 1562–1566 (2003).
3. Ramelot, T. A. *et al.* Solution structure of the yeast ubiquitin-like modifier protein Hub1. *J. Struct. Funct. Genomics* **4,** 25–30 (2003).
4. Wilkinson, C. R. *et al.* Ubiquitin-like protein Hub1 is required for pre-mRNA splicing and localization of an essential splicing factor in fission yeast. *Curr. Biol.* **14,** 2283–2288 (2004).
5. Friedman, J. S., Koop, B. F., Raymond, V. & Walter, M. A. Isolation of a ubiquitin-like (*UBL5*) gene from a screen identifying highly expressed and conserved iris genes. *Genomics* **71,** 252–255 (2001).
6. Dittmar, G. A., Wilkinson, C. R., Jedrzejewski, P. T. & Finley, D. Role of a ubiquitin-like modification in polarized morphogenesis. *Science* **295,** 2442–2446 (2002).
7. Lüders, J., Pyrowolakis, G. & Jentsch, S. The ubiquitin-like protein HUB1 forms SDS-resistant complexes with cellular proteins in the absence of ATP. *EMBO Rep.* **4,** 1169–1174 (2003).
8. Yashiroda, H. & Tanaka, K. Hub1 is an essential ubiquitin-like protein without functioning as a typical modifier in fission yeast. *Genes Cells* **9,** 1189–1197 (2004).
9. Benedetti, C., Haynes, C. M., Yang, Y., Harding, H. P. & Ron, D. Ubiquitin-like protein 5 positively regulates chaperone gene expression in the mitochondrial unfolded protein response. *Genetics* **174,** 229–239 (2006).
10. Hazbun, T. R. *et al.* Assigning function to yeast proteins by integration of technologies. *Mol. Cell* **12,** 1353–1365 (2003).
11. Wahl, M. C., Will, C. L. & Luhrmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136,** 701–718 (2009).
12. Ramage, R. *et al.* Synthetic, structural and biological studies of the ubiquitin system: the total chemical synthesis of ubiquitin. *Biochem. J.* **299,** 151–158 (1994).
13. Bayer, P. *et al.* Structure determination of the small ubiquitin-related modifier SUMO-1. *J. Mol. Biol.* **280,** 275–286 (1998).
14. Dikic, I., Wakatsuki, S. & Walters, K. J. Ubiquitin-binding domains—from structures to functions. *Nature Rev. Mol. Cell Biol.* **10,** 659–671 (2009).
15. Song, J., Zhang, Z., Hu, W. & Chen, Y. Small ubiquitin-like modifier (SUMO) recognition of a SUMO binding motif: a reversal of the bound orientation. *J. Biol. Chem.* **280,** 40122–40129 (2005).
16. Makarova, O. V., Makarov, E. M. & Luhrmann, R. The 65 and 110 kDa SR-related proteins of the U4/U6.U5 tri-snRNP are essential for the assembly of mature spliceosomes. *EMBO J.* **20,** 2553–2563 (2001).
17. Stevens, S. W. & Abelson, J. Purification of the yeast U4/U6.U5 small nuclear ribonucleoprotein particle and identification of its proteins. *Proc. Natl Acad. Sci. USA* **96,** 7226–7231 (1999).
18. Stevens, S. W. *et al.* Composition and functional characterization of the yeast spliceosomal penta-snRNP. *Mol. Cell* **9,** 31–44 (2002).
19. Fabrizio, P. *et al.* The evolutionarily conserved core design of the catalytic activation step of the yeast spliceosome. *Mol. Cell* **36,** 593–608 (2009).
20. Clark, T. A., Sugnet, C. W. & Ares, M. Jr. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* **296,** 907–910 (2002).
21. Burckin, T. *et al.* Exploring functional relationships between components of the gene expression machinery. *Nature Struct. Mol. Biol.* **12,** 175–182 (2005).
22. Keren, H., Lev-Maor, G. & Ast, G. Alternative splicing and evolution: diversification, exon definition and function. *Nature Rev. Genet.* **11,** 345–355 (2010).
23. Jacquier, A., Rodriguez, J. R. & Rosbash, M. A quantitative analysis of the effects of 5′ junction and TACTAAC box mutants and mutant combinations on yeast mRNA splicing. *Cell* **43,** 423–430 (1985).
24. Lesser, C. F. & Guthrie, C. Mutational analysis of pre-mRNA splicing in *Saccharomyces cerevisiae* using a sensitive new reporter gene, *CUP1*. *Genetics* **133,** 851–863 (1993).
25. Chen, M. & Manley, J. L. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature Rev. Mol. Cell Biol.* **10,** 741–754 (2009).
26. Davis, C. A., Grate, L., Spingola, M. & Ares, M. Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. *Nucleic Acids Res.* **28,** 1700–1706 (2000).
27. Rodríguez-Navarro, S., Igual, J. C. & Pérez-Ortín, J. E. *SRC1*: an intron-containing yeast gene involved in sister chromatid segregation. *Yeast* **19,** 43–54 (2002).
28. Grund, S. E. *et al.* The inner nuclear membrane protein Src1 associates with subtelomeric genes and alters their regulated gene expression. *J. Cell Biol.* **182,** 897–910 (2008).
29. Kuhn, A. N. & Kaufer, N. F. Pre-mRNA splicing in *Schizosaccharomyces pombe*: regulatory role of a kinase conserved from fission yeast to mammals. *Curr. Genet.* **42,** 241–251 (2003).
30. Okazaki, K. & Niwa, O. mRNAs encoding zinc finger protein isoforms are expressed by alternative splicing of an in-frame intron in fission yeast. *DNA Res.* **7,** 27–30 (2000).
31. Kress, T. L., Krogan, N. J. & Guthrie, C. A single SR-like protein, Npl3, promotes pre-mRNA splicing in budding yeast. *Mol. Cell* **32,** 727–734 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank U. Cramer and M. Kost for technical assistance, G. Dittmar, C. Guthrie, M. Konarska, K. Matuschewski, O. Nielsen, M. Rosbash and H. Yashiroda for materials, S. Uebel and C. Boulegue for mass spectrometric analysis and help, K. Hofmann for pointing out putative HIND elements in Prp38 proteins of plants, and M. Singh for initiating structural work. S.J. is supported by the Max Planck Society, Deutsche Forschungsgemeinschaft, Fonds der chemischen Industrie, Center for Integrated Protein Science Munich and RUBICON EU Network of Excellence; T.A.H. by the Max Planck Society; M.A. by NIH (GM040478).

Author Contributions S.K.M. (*S. cerevisiae, S. pombe*), T.A. (mammalian) and S.J. designed, obtained and analysed the genetic, biochemical and functional data; G.M.P, M.K. and T.A.H. the structural data; R.J.N. and M.A. Jr the splicing array data. S.J. and S.K.M. wrote the paper, and all authors contributed to the manuscript.

Author Information Coordinates and the experimental structural factors of both complexes have been deposited in the PDB under the following codes: Hub1–HIND-I, 3PLU; Hub1–HIND-II, 3PLV. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.J. (Jentsch@biochem.mpg.de).

METHODS

Yeast strains, plasmids and DNA techniques. *S. cerevisiae* and *S. pombe* strains are listed in Supplementary Table 1. Yeast growth assays were performed by spotting fivefold serial dilutions of indicated strains on solid agar plates. Yeast strains isogenic to DF5, W303 and BY4741 were used for biochemical and genetic studies, and PJ69-7a for Y2H assays. All tagged strains and deletion mutants were constructed by a PCR-based strategy[32,33] and confirmed by western blot analysis and PCR, respectively. The *prp8** allele (*prp8(P1384L)*) was provided by G. Dittmar. For complementation studies, the complete *HUB1* and *SNU66* open reading frame (ORF), 1,000 bp of the upstream promoter and 500 bp of its terminator were cloned into centromeric plasmids. For Hub1 overexpression, the *HUB1* ORF was cloned into an integrative plasmid with the *TEF2* promoter and *ADH1* terminator. To generate 5′-splice-site variants for the *SRC1* gene, the full-length ORF was cloned into a centromeric plasmid downstream of the *GAL1-10* promoter. To generate the Hub1–Snu66ΔN linear fusion, DNA sequences encoding Hub1 plus a linker of nine glycine residues was inserted downstream of the *SNU66* promoter and upstream of the *SNU66* ORF. The vectors *pET28a-c* or *pGEX-5X1* were used for expression of 6×His-tagged or GST-tagged proteins, respectively, and *pGBDUC1* and *pGADC1* for Y2H assays. Plasmids with point mutations were constructed by site-directed mutagenesis using specific primers. *Plasmodium falciparum* genomic DNA (gift from K. Matuschewski) or a cDNA library of *Arabidopsis thaliana* (BioChain) were used as templates to generate the corresponding clones for interaction studies. For western-blot-based *SRC1* alternative splicing assays, a DNA sequence covering the *CUP1-1* promoter and TAP tag was integrated upstream of *SRC1* at its genomic loci. The *CUP1-1* promoter was induced for 4 h in the presence of 200 μM $CuSO_4$. For the construction of linear fusions of the splicing factors Snu66, Prp38 and Prp8 with Hub1 variants, DNA encoding *HUB1*, *hub1(D22A)* and *hub1–DD* were cloned into vector *pFA6a-natNT2*. The C-terminal fusions were finally generated by a PCR-based strategy[32,33]. For the (*GAL1-10*-promoter-driven) splicing reporter assays (Fig. 3a–c) and *SRC1* alternative splicing assays (Supplementary Fig. 13d), cells were grown to log phase in synthetic media containing 2% lactate as carbon source (pH 5.5). Expression by the *GAL1-10* promoter was induced for the indicated time by adding 2% galactose. The promoter was shut off by adding 2% glucose.

S. pombe strains are isogenic to JY741/746. *S. pombe* strains *Δhub1* (YHY23P) and *hub1-1* (YHY24P), and plasmids *pUR19 hub1+* and *pALSK* are gifts from H. Yashiroda[8]. Expression vectors for *S. pombe* studies were purchased from ATCC. For *S. pombe Δhub1* complementation assays, *HUB1* cDNA was cloned into *pREP81* (a vector containing thiamine-repressible *nmt81* promoter). For 5-FOA shuffle and complementation assays of *S. pombe Δsnu66*, the *SNU66* ORF with 1,000 bp of the upstream promoter and 500 bp of terminator were cloned into vectors *pUR19* (*ura4+*) or *pALSK* (*leu2+*). For complementation assays in *S. pombe*, competent cells were transformed with plasmids and incubated on selective media for 4 days at 30 °C. Fivefold serial dilutions were spotted on control and 5-FOA selection plates. The plates were incubated for 3–5 days at the temperatures indicated. Immunofluorescence in *S. pombe* was performed following a published protocol[34].

Antibodies. The following antibodies were used: anti-haemagglutinin (HA, clone F-7) and anti-Clb2 (Santa Cruz Biotechnology); anti-Myc (clone 9E10), anti-Flag (clone M2) and anti-TAP (PAP) (all from Sigma-Aldrich); sheep anti-mouse Cy3-conjugated and HRP-coupled secondary antibodies (Jackson Immuno Research); antibodies against purified recombinant Hub1 (raised in rabbit); antibodies raised against two different Snu66 peptides (Eurogentec).

Protein techniques and interaction studies. Standard procedures were followed for purification of GST- and 6×His-tagged recombinant proteins from *Escherichia coli*. Proteins were dialysed and then cleaned on size exclusion columns. Because Snu66 tends to precipitate during dialysis, the salt concentration was increased to 300 mM NaCl. Na-phosphate-based buffers were used for interaction studies, whereas HEPES-based buffer was used for proteins for structural studies. Isothermal titration calorimetry was performed to measure dissociation constants (K_d) using purified proteins and peptides. For Y2H assays, strain PJ69-7a was co-transformed with plasmids expressing the Gal4 DNA-binding domain fused to the baits, and the Gal4 activation domain fused to the preys. After 3 days of growth on selective media at 30 °C, fivefold dilutions of cultures were spotted on control plates and plates lacking the indicators histidine (-his) or adenine (-ade). The plates were further incubated for 2–5 days at 30 °C. For co-immunoprecipitation assays, 100 OD_{600} yeast cells were harvested at an $OD_{600} = 2.0$. Cells were washed once with PBS, and the pellet was frozen in liquid nitrogen. Cells were lysed either by bead beating (Retsch Instrument) or grinding in liquid nitrogen. The total yeast lysate prepared in a PBS-based buffer (1% Triton X-100, protease inhibitors by Roche), was subjected to immunoprecipitation with bead-coupled antibodies for 3 h at 4 °C. Beads were washed 4 times by rotating for 5 min at 4 °C in 1 ml lysis buffer. Bound material was eluted by a dithiothreitol-containing hydroxyurea buffer[32] followed by SDS–PAGE (4–16%) and western blot analysis. For *in vitro* GST pull-down assays, 50 μg of GST-fusion proteins bound to glutathione Sepharose beads were used to pull-down from equimolar amounts of purified proteins using the same buffer conditions as for co-immunoprecipitation experiments. Pull-down was performed for 1 h at 4 °C; subsequent washing of beads, elution of bound material and electrophoresis was similar as for the co-immunoprecipitation assay. Protein bands were stained with PageBlue (Fermentas). For HIND pull-down assays, purified Snu66-N (wild type or RR-AA mutant) was covalently coupled to Sepharose beads (GE Healthcare). Pull-down of interacting proteins from yeast lysate was performed similar to the co-immunoprecipitation experiments. For immunoprecipitations using mammalian proteins, transiently transfected 293T cells were harvested, washed in ice-cold PBS and cell pellets were lysed at 4 °C for 30 min in 5 pellet volumes of lysis buffer (50 mM HEPES pH 7.2, 150 mM NaCl, 2 mM EDTA, 1% Triton X-100, 1 mM PMSF, and complete protease inhibitors (Roche)). After removal of cell debris by centrifugation (10 min, 16,000*g*, 4 °C), lysates were incubated with anti-Flag-M2 affinity gel (Sigma-Aldrich). After 3 h the affinity matrix was washed 5 times in lysis buffer and eluted in Laemmli SDS buffer.

Purification of Snu66-associated complexes. For each purification (immunoprecipitation) reaction, 200 μg anti-Snu66 antibody coupled to Dynal magnetic beads (Invitrogen) was used. Yeast lysate (~12 ml) was prepared from 3,500 OD_{600} of yeast pellet by grinding with liquid nitrogen as described[35]. Immunoprecipitation was performed at 4 °C for 3 h. Beads were washed 4 times with 12 ml lysis buffer containing 1% Triton X-100 and rotating for 5 min at 4 °C. Co-immunoprecipitation material was eluted by heating the beads at 65 °C in the presence of 2% SDS then separated on 4–12% SDS–PAGE. Protein bands were stained with PageBlue (Fermentas). Each sample lane was cut into nine gel slices, proteins were extracted, digested with trypsin, and analysed by Orbitrap mass spectrometry[36].

Splicing-specific microarrays. Splicing-specific microarrays, experimental design and data processing were described previously[20,21].

Northern blot analysis and RT–qPCR. RNA was isolated for northern blot analysis using TRIzol (Invitrogen) and for RT–qPCR using an RNeasy kit (Quiagen). ^{32}P-labelled *HUB1* and *SMT3* (SUMO)-specific probes were synthesized by random primer labelling (Ambion). Light cycler 480 was used for RT–qPCR assays (Roche). Fifteen micrograms of total RNA was used for northern blot assay whereas 500 ng of total RNA was used for reverse transcription.

Splicing reporter assays. The β-galactosidase assays were performed as described[23]. Site-directed mutagenesis was used to generate 5′-splice-site variants of the *RP51–LacZ* construct (gift from M. Rosbash). Survival on $CuSO_4$-containing solid media for *ACT1–CUP1* fusions was performed as described[24]. Yeast strain yJU75, plasmids bearing *PRP8* and *prp8-101*, and *ACT1–CUP1* reporters wild type and UuG were gifts from C. Guthrie; plasmids *prp8-R1753K*, *prp8-161* and *prp8-162*, and *ACT–CUP1* reporters A3C, BS-C and BS-G were gifts from M. Konarska.

Human cell lines, transfections and clones. HEK 293T and U2OS cells were maintained in DMEM (PAA) supplemented with 10% FCS (Biochrom AG) at 37 °C, 5% CO_2. HEK 293T cells were transfected using the calcium phosphate precipitation technique. Transfection of U2OS was performed using Lipofectamine 2000 (Invitrogen). The cDNA clone for Hs Snu66 (SART1) was purchased from Origene. The cDNA for Hs Hub1 (UBL5) was amplified by RT–PCR using total RNA from HeLa cells. Standard cloning techniques were used to generate expression constructs in *pEGFP-N1* (Clontech) or *p3×Flag-CMV-10* (Sigma-Aldrich) vectors.

Immunofluorescence microscopy. For immunofluorescence microscopy U2OS cells were seeded and transfected on glass coverslips (Roth). Cells were washed twice with PBS and fixed in 3.7% fresh paraformaldehyde (18 min, room temperature (24 °C)). After incubation, paraformaldehyde was removed by aspiration and the cells were washed three times in PBS (5 min each). Permeabilization of cells was performed using PBS-Triton X-100 (0.4% for 6 min), followed by three PBS-Tween (Tween 0.05%; PBS-T) washing steps and blocking in PBS-T/2% BSA for 1 h at room temperature. Coverslips were incubated with primary antibody for 2 h (dilution 1:200 in blocking buffer) and then washed three times in PBS-T (3 min each, room temperature). After incubation with secondary antibody (sheep anti-mouse Cy3-conjugated, Jackson Immuno Research), cover slips were mounted using DAPI-containing mounting medium (Vectashield, Vector Labs). Images were acquired on a Zeiss AxioImager Z1 microscope equipped with an AxioCam MRm Rev.3 camera. Image acquisition was carried out using AxioVision Rel. 4.7 software (Zeiss).

X-ray crystallography. Complexes of recombinant 6×His-tagged Hub1 and HIND peptides were purified on S75 Superdex column. The protein buffer used for crystallization contained 20 mM HEPES and 100 mM NaCl (pH 7.4). Crystallization of the complexes was carried out with the sitting-drop vapour diffusion method at 20 °C by mixing equal volumes of protein complex and

reservoir solution (0.2 M ammonium iodide, 20% PEG 3350, pH 6.9). Crystals of both complexes appeared in several weeks and grew to a final size of ~0.3 × 0.2 × 0.1 mm. Crystals were plunged frozen after soaking for ~30 s in a drop of a reservoir solution containing 30% v/v glycerol as a cryoprotectant. The crystals of the HIND-I–Hub1 complex belong to the space group $P1$, with the unit cell $a = 35.2$ Å, $b = 36.34$ Å, $c = 36.78$ Å, $\alpha = 83.44^\circ$, $\beta = 89.85^\circ$, $\gamma = 85.84^\circ$ and contained two complexes per an asymmetric unit. The HIND-II–Hub1 complex crystallized in space group $P2_12_12$, with the unit cell dimensions $a = 36.72$ Å, $b = 83.57$ Å, $c = 35.11$ Å, and contained one complex per asymmetric unit.

The data sets, up to 1.4 Å and 1.9 Å at 90 K, were collected on the MPG/GBF beamline BW6 at DESY, Hamburg, Germany using 1.05 Å wavelength. The collected data were integrated, scaled and merged by XDS and XSCALE programs[37]. The structure was determined by molecular replacement using the Molrep program from the CCP4 suite[38]. The structure of ubiquitin, taken from the PDB entry 1UBI, was used as a probe structure. Models were then refined by Refmac5 (ref. 38) and rebuilt by XtalViev/Xfit[39], followed by a subsequent Refmac5 refinement. Waters were added by Arp/warp[40]. The complete Hub1 molecule model had clear interpretable electron density, except for certain solvent exposed side chains, and those parts were omitted in the model. The structure of the Hub1–HIND-I complex has 94.9% of residues in the core, and 5.1% in allowed Ramachandran regions. For the Hub1–HIND-II complex those values are 93% and 7% respectively. Data collection and refinement statistics are summarized in Supplementary Table 2.

NMR spectroscopy. NMR spectroscopy was performed on a BRUKER AVANCE 600 MHz spectrometer equipped with a Cryo-probe. For each sample a 1H-1D spectrum with a WATERGATE-5 water suppression was measured, and for 15N labelled samples a fast HSQC spectrum[41,42] was recorded. For the full-length 2H-15N Snu66 experiment also a Trosy-HSQC was recorded[43,44].

CD spectroscopy. Synthetic peptides were solubilised in a buffer containing 20 mM NaH_2PO_4, 100 mM NaCl (pH 7.6) at 0.01–0.02 mM concentration. CD spectra were obtained on a J-715 spectropolarimeter (Jasco J715 model). All spectra were recorded by using a quartz cell with a path length of 1 mm. The parameters used for data acquisition were: response, 2 s; scanning speed, 20 nm min^{-1}; bandwidth, 1.0 nm; sensitivity, 5 mdeg; and step resolution, 0.1 nm.

32. Knop, M. *et al.* Epitope tagging of yeast genes using a PCR-based strategy: more tags and improved practical routines. *Yeast* **15,** 963–972 (1999).
33. Janke, C. *et al.* A versatile toolbox for PCR-based tagging of yeast genes: new fluorescent proteins, more markers and promoter substitution cassettes. *Yeast* **21,** 947–962 (2004).
34. Hagan, I. M. & Ayscough, K. R. in *Protein Localization by Fluorescence Microscopy: A Practical Approach* (ed Allan, V. J.) 179–205 (Oxford Univ. Press, 2000).
35. Ansari, A. & Schwer, B. SLU7 and a novel activity, SSF1, act during the PRP16-dependent step of yeast pre-mRNA splicing. *EMBO J.* **14,** 4001–4009 (1995).
36. Steen, H. & Mann, M. The abc's (and xyz's) of peptide sequencing. *Nature Rev. Mol. Cell Biol.* **5,** 699–711 (2004).
37. Kabsch, W. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Cryst.* **26,** 795–800 (1993).
38. Collaborative Computational Project 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
39. McRee, D. E. XtalView/Xfit—a versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **125,** 156–165 (1999).
40. Lamzin, V. S. Automated refinement of protein models. *Acta Crystallogr. D* **49,** 129–147 (1993).
41. Liu, M. *et al.* Improved WATERGATE pulse sequences for solvent suppresion in NMR spectroscopy. *J. Magn. Reson.* **132,** 125–129 (1998).
42. Mori, S., Abeygunawardana, C., Johnson, M. O. & van Zijl, P. C. Improved sensitivity of HSQC spectra of exchanging protons at short interscan delays using a new fast HSQC (FHSQC) detection scheme that avoids water saturation. *J. Magn. Reson. B.* **108,** 94–98 (1995).
43. Zhu, G., Kong, X. M. & Sze, K. H. Gradient and sensitivity enhancement of 2D TROSY with water flip-back, 3D NOESY-TROSY and TOCSY-TROSY experiments. *J. Biomol. NMR* **13,** 77–81 (1999).
44. Pervushin, K. V., Wider, G. & Wüthrich, K. Single transition-to-single transition polarization transfer (ST2-PT) in [15N,1H]-TROSY. *J. Biomol. NMR* **12,** 345–348 (1998).